Demystifying the Semantics of Relevant Objects in Scholarly Collections: A Probabilistic Approach

José María González Pinto
IFIS TU Braunschweig
Mühlenpfordstrasse 23
38106 Braunschweig, Germany
+49 (531) 391 2232
pinto@ifis.cs.tu-bs.de

Wolf-Tilo Balke
IFIS TU Braunschweig
Mühlenpfordstrasse 23
38106 Braunschweig, Germany
+49 (531) 391 3271
balke@ifis.cs.tu-bs.de

ABSTRACT

Efforts to make highly specialized knowledge accessible through scientific digital libraries need to go beyond mere bibliographic metadata, since here information search is mostly entity-centric. Previous work has realized this trend and developed different methods to recognize and (to some degree even automatically) annotate several important types of entities: genes and proteins, chemical structures and molecules, or drug names to name but a few. Moreover, such entities are often crossreferenced with entries in curated databases. However, several questions still remain to be answered: Given a scientific discipline what are the important entities? How can they be automatically identified? Are really all of them relevant, i.e. do all of them carry deeper semantics for assessing a publication? How can they be represented, described, and subsequently annotated? How can they be used for search tasks? In this work we focus on answering some of these questions. We claim that to bring the use of scientific digital libraries to the next level we must find treat topic-specific entities as first class citizens and deeply integrate their semantics into the search process. To support this we propose a novel probabilistic approach that not only successfully provides a solution to the integration problem, but also demonstrates how to leverage the knowledge encoded in entities and provide insights to explore the use of our approach in different scenarios. Finally, we show how our results can benefit information providers.

Categories and Subject Descriptors

H.3.7 [INFORMATION STORAGE AND RETRIEVAL]: Digital Libraries

General Terms

Experimentation, Algorithms, Reliability.

Keywords

Scientific Digital Libraries, hidden knowledge, semantics entities, probabilistic topic models.

JCDL'15, June 21 - 25, 2015, Knoxville, TN, USA
© 2015 ACM. ISBN 978-1-4503-3594-2/15/06…$15.00
DOI: http://dx.doi.org/10.1145/2756406.2756923

1. INTRODUCTION

A lot of effort in the Digital Library community is put on assuring high quality metadata, because it is considered vital for search tasks. Yet most of this metadata is merely of bibliographic nature like title, author, or year of publication. In contrast, search tasks often are *entity-centric* as has been shown especially in scientific digital libraries: For instance, in the biomedical domain special tools for identifying and extracting genes, proteins, and drug names [4, 5] have been developed. In chemistry relevant entities, e.g., chemical substances, mulecules, and reactions, have yielded commercial tools for automatic recognition such as CliDE[1] together with focused research efforts such as [6–8] and of course as a human effort-based provider: the Chemical Abstract Service (CAS[2]). In physics the manually curated portal ScienceWISE[3] offers entity-based access to information, including also a variety of mathematical expressions, too. And also for such formulae we find lots of tools such as Symbolab[4], Springer LATEX[5] and research efforts [9–11] in the information retrieval community.

Indeed, the relevance of entities for scientists is paramount. But the challenge for indexing entities in a meaningful way is to really capture their semantics, including the role of the entities [11–13] in each document. Certainly this is a major problem still needing attention from the research community: *How to semantically represent or at least annotate those entities found in documents that cannot be indexed by current state of the art indexing mechanisms?* After all, considering chemical molecules, gene sequences, mathematical formulae, etc. What are these entities? How to describe them? What do they mean? How can we validate the relevance of these entities regarding a document?

To make our point clear, consider the classic representation of the well-known Pythagorean theorem: $a^2 + b^2 = c^2$ and imagine a student looking for papers in a Scientific Digital Library where this theorem is actually used. First our student might type in the equation to find a set of relevant documents. Now, to what degree could e.g., a document containing the equation $x^2 + y^2 = z^2$ be considered to fulfill the student's information need? Currently there are several research projects underway trying to figure out

[1] http://www.simbiosys.ca/clide/

[2] http://www.cas.org/

[3] http://www.sciencewise.info/

[4] http://www.symbolab.com/

[5] http://www.latexsearch.com/

structural similarities between mathematical expressions, some go down to trying to fully understand mathematical equivalence of even complex expressions. But is such a degree of effort really the only solution to sufficiently capture entity semantics?

A viable alternative to answer the question might be to only look at the *surrounding context* of the equation and then try to index the entity in a conventional fashion. While this often proves to work well, we are still facing two challenges:

- Among the variety of instances of the same class of entities we need to determine a criterion to identify those, which are suitable in terms of relevance and useful for a given task;
- We need to find a suitable representation of each entity found according to the above criterion, aiming at uncovering its semantics.

In this paper we tackle the problem of capturing relevant entity semantics for publications and implement a practical model to meet both challenges: to discover the set of relevant entities from some collection, and to encode these entities' semantics. We also validate the indexed entities' usefulness in a real world use case.

To discover the set of relevant entities we propose a probabilistic approach to perform an analysis of the collection and deal effectively with the uncertainty of the selection of relevant entities. In our setting, each document –after removing non-entity elements is seen as a bag of entities. On this representation we apply a probabilistic topic model over the entities. Once we have selected all relevant entities from the probabilistic model, we proceed to discover their semantics. We propose an innovative application of a probabilistic topic model to capture the different intended meanings of each of the entities. Our intuition in this second model is based on the assumption that "You shall know a word by the company it keeps . . ." [18]. In our work we argue that by capturing the companionship of each entity, we can generate a highly effective approximation of the entities' meaning and hence their semantics. Finally, we use an application oriented approach to validate the relevance of the entity semantics. In summary, our contributions are as follows:

- We design and implement a model to find relevant entities and represent and its semantics in a collection.
- We validate our model with a case study from mathematics:
    - We harness a highly specialized collection to find relevant entities (formulae).
    - We discover through several experiments how to model the semantics of the entities.
    - We provide a sample application case to validate our model in a very difficult but interesting task: given a document predict its corresponding taxonomy main class.
- We yield evidence of the validity of our approach to apply it in other domains.

The rest of the paper is organized as follows: in section 2 we will give an overview of related work. Section 3 introduces the model and the problem formulation. In section 4 we present our case study and results. And in section 5 we conclude with a summary and future work.

2. BACKGROUND AND RELATED WORK

Our approach builds on ideas from the Natural Language Processing (NLP) community. In particular, from the task of word sense induction. The idea is the following: any representative object –as words can have different meanings depending on its surrounded context. In NLP the intuition is that it is possible to infer the sense of a word automatically using unsupervised learning methods. Previous work in machine translation [14, 15] gave evidence that supports the idea and inspire us to design a model base on the results found in the literature. Furthermore, the work of [16, 17] explores the use of topic models for a very similar challenge but here we are dealing with a entirely different scenario –entities. Indeed, here instead of looking at words, where exists belief of the number of senses, we look at entities with the similar underlying assumptions.

For instance, given any entity of interest, we can imagine that it can have different meanings depending on its use. How do we know the intended meaning? The answer is straightforward: by its context. Therefore, identifying and modeling properly the context is the key aspect to consider to solve the problem. But finding the correct context is not trivial and it depends on the data. Relying on experts is not an option because data grows exponentially and it would be not only too expensive but also too much time consuming to solve the problem. Thus, we explore the use of Latent Dirichlet Allocation (LDA)[1] as our main data driven approach for this challenge. Probabilistic topic models have been used and adapted in different kinds of data and applications: to find patterns in genetic data, social networks, word sense induction and disambiguation, etc. These algorithms have also the advantage that they can be applied to massive collections of documents [3].

In the interest of generating useful context we implement several models to capture the intended meaning of an entity. Our hypothesis is that once we can find the proper context, we cannot only use these objects as features for different types of applications: metadata annotation, classification, cluster analysis, etc. but also validates their use and relevance for a given task.

3. MODEL AND PROBLEM FORMULATION

Here, we argue that we can find and represent any type of objects with its semantics from a collection. Concretely:
Given a set of documents $C = \{d_1, \dots, d_n\}$ and a definition of a class of representative entities –gene sequences, chemical molecules, mathematical, etc., the task can be summarized in the following steps:

1. Select a relevant set of objects from the collection.
2. Find the semantics of the set of relevant objects.

In order to apply these two steps, the collection C has to be preprocessed and transformed. For the first step, the document collection is modeled as a "bag of entities". So we do as sketched in Algorithm 1. Once we have performed this process –depicted in Figure 1, we have our collection $C_{entities}$ which contains the same number of documents of the original collection but only with the entities as content. In section 3.1 we will refer to this collection for the process of objects' selection. For our second step, we describe in section 3.2 our pre-processing of the collection C, given the output of our selection step.

```
Algorithm 1 Generate collection of entities
Input: the collection $C$
Ouput: the collection $C_{entities}$

for each document $d$ in the collection $C$ do
  create a new empty document $d_{entities}$
  extract from $d$ all the entities
  assign extracted entities to $d_{entities}$
end for
```

We use LDA in the two steps. Here we give the intuition behind our approach. Basically, LDA captures the idea of learning from data without the need of coercing a document to be only about one single topic. Instead, the intuition is to talk about a hidden variable model of documents. In LDA, what we observed are the words of each document and the hidden variables represent the latent topical structure: the topics and the way each document exhibits them [2]. Figure 2 illustrates the latent Dirichlet allocation model.

More formally what we have in a typical LDA is the following:

1. For each topic,
   a. Draw a distribution over words β_k
2. For each document,
   a. Draw a vector of topic proportions θ_d
   b. For each word,
      (a) Draw a topic assignment $Z_{d,n}$

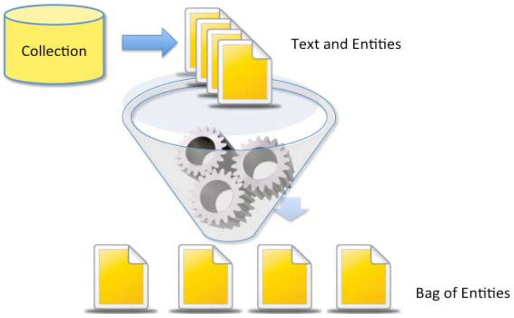

Figure 1. From a collection of text and entities to a collection of entities.

The distribution over the words given the number of k topics is assumed to be a Dirichlet as well as the distribution over the topic proportions. In the first step of our model the observed variable is the entity and it is what we called the LDA over instances.
Certainly, in our first step the assumption is the notion of the "bag of entities" instead of the traditional view of the "bag of words" in text processing scenarios in which probabilistic topic models have been extensively used.

And for our second step our observed variable are the context window words which captures the semantics of each entity. Indeed in this second step we put the model in a situation in which each of the documents to use in the process is rather small. In this second step therefore, the number of topics gives the number of senses for each entity. In the following subsections we give the details of each of these two steps and we use object and instance interchangeable to refer to a specific example of the class of entities we are interested.

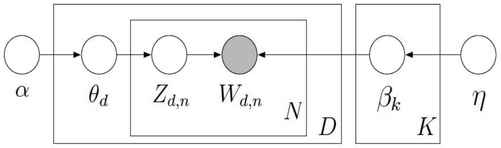

Figure 2. Graphical model representation of LDA.

3.1 Object selection

First of all, our model assumes that exists a way to find and extract from each document in the collection every instance of an entity. Thus, this information is available to us by some preprocessing step over the collection –a named entity recognition tool. Our problema reduces to extract a relevant set of entities from these objects. Our approach relies on LDA to find the set of relevant instances, which can be seen as extracting the most probable instances from the most probable topics from the collection. To perform this step, we use the collection $C_{entities}$ described in the heading of this section. This is what we call the LDA over instances. After some empirical investigation we set the number of topics t and the parameters α and β to obtain the set of representative objects. For our sample case, we show in section 4 the dataset and the heuristics used for selection. Notice that it is possible to go to the second step of our model if we already have the set of objects to work with. However, we provide a full solution to the problem in cases where there is no such a set and hence has to be infer from data.

3.2 Semantics of the set of relevant objects

Given the set of relevant objects, we proceed with the next step of our approach: find the semantics of each of the instances. The intuition of our approach considers the following: every object that occurs t times in the collection has s senses –semantic meaning. The semantic meaning of an object is given by its context, in our case the words around the instance. The challenge for us is to find the correct number of senses and the correct number of words as context. First, we perform over the collection lower case transformation, we remove punctuation and stop words. We keep the instances of interest in its original form in the document. The modified collection with this preprocessing steps is what we call C_{trans}. Once we have created this new curated collection, we proceed as follows:

Given a set of objects $O = \{o_1, \ldots, o_n\}$ obtained from section 3.1 and a collection of documents $C_{trans} = \{d_1, \ldots, d_n\}$:

```
1. For each document $d$:
      (a) For each $o$ found in $d$
      i. Extract context sizes $s$ window from $d$ and generate a new document $d'_{od}$
      ii. Add the $d'_{od}$ to the collection of $Context_{objects}$
2. Save the new collection $Context_{objects}$ which has context size $s$.
```

With this new collection we investigate the derived number of likely senses –semantics for each entity from the set of relevant

Collection of a given context window size + Set of entities to represent
Subcollections capturing the semantics of each entity

Figure 3. Process to find the semantics of the relevant entities.

entities as outlined below:

1. For each object in $O = \{o_1, \dots, o_n\}$
   (a) Select from collection Context$_{\textbf{objects}}$ the documents which correspond to object o
   (b) Apply LDA in this sub-collection to infer the semantics of object o

Figure 3 shows the idea behind this step. We will be generating n LDA models for each context window size collection, where n is the number of entities selected in the first step. Our model for this step then has two additional parameters: s –the size window and n–number of senses. Since we have neither any previous knowledge of the correct number of senses nor of the correct number for the size context parameter we will perform step 2 with different values of n and s.

Notice that in the Context$_{\text{objects}}$ collection each document is small and contains the words-context window of a particular object appearing in a particular document. In other words, in each d´$_{\text{od}}$ the subindex stands for an object o found in document d of the collection C$_{\text{entities}}$. This representation allow us to find the semantics of each of the objects given that we know which files in the collection C$_{\text{objects}}$ accounts for each object o.

In section 4 we present a use case to show how to deal with the parameters s and n.

4. EXPERIMENTS

For our approach to work, there must have a predefined annotated entities in a given collection. One can for instance use a Named Entity Recognition tool and apply it to the collection or use a collection with this preprocessing step already performed by such tools. The second thing that we need is a way to extract the text from the document collection. This can be done with a tool such as Apache Tika[6] or a similar framework. Once these two prerequisites have been satisfied we can start with the model.

We selected mathematical expressions as our case study due to the ease of identifying them in an open collection that was provided to us. Formulae have received attention from the information retrieval community but from the perspective of encoding its structure using different approaches and heuristics. The main idea behind these attempts is to index formulae using traditional methods of the information retrieval community with some variants splitting formulae into tokens. In our case study, we turn our interest in the analysis of the semantics of the formulae. Therefore, instead of decomposing the formulae in tokens we take a formula as an entity –a relevant object of a document and discover a way to represent its semantics. Our case study thus complements math information retrieval's previous attempts by providing plausible ways to attach keywords to a formula capturing its use and, therefore, its intended meaning.

4.1 Data set description

We use a dataset of 101,120 XHTML5 files from the collection of the NTCIR-11 Math Retrieval Task. The collection contains documents from Mathematical Physics, High Energy Physics, Computer Science, Nonlinear Sciences and Statistics. However, for our application guided approach –the classfication task, we needed for each document in the collection its corresponding Mathematics Subject Classification (MSC[7]). The MSC is a taxonomy used by several mathematical journals to help users find the documents of potential interest to them. And since the collection does not contain the MSC classes of the documents we had to do some pre-processing to gather them from the source Cornell University Library[8]. After harvesting the collection to account for the MSC classes of each document of the collection we end up with 56,051 files. This accounts for a total of 37,075,959 mathematical expressions, including monomial expressions.

4.2 Model and set of relevant objects selection

In our first step we need to find the number of relevant object by applying LDA over the collection $C_{entities}$. Remember that in this model we use the notion of a bag of entities, mathematical expressions in out case study. The model has the following parameters: the number of topics, α, β and the number of iterations for the Gibbs sampler. For the first challenge –selecting the number of topics, we first need to reduce the vocabulary size. Usually one has at least two options to reduce the vocabulary size: selecting only terms –entities in our case that occur in a minimum number of documents [19] or using the highest term-frequency inverse document frequency (tf-idf) scores such as in [2] We use these successful experiences with words as our guidance for our first step. We opted to use the tf-idf score to select the vocabulary only, but we use the frequencies of the remaining entities to feed the LDA. Because we are dealing with entities, we could not use any other preprocessing over our collection such as lower case conversion, stemming and omitting terms below a certain minimum length.

Indeed, we are considering each entity as a black box; in other words, we do not perform any modification of the manner the entities were written in the collection. We did remove punctuation characters and through empirical exploration we create our stop entities list –similar to the notion of stop words list using the tf-idf score. For our second challenge, selecting the number of topics, we started by using the number of different MSC top clases represented in our collection and fixing α and β to investigate the effect of varying the number of topics from the initial number and up to 400 which seems to work in practice given the size of the collection. We measure this effect by selecting from each model

[6] http://tika.apache.org/

[7] http://www.ams.org/msc/msc2010.html

[8] http://arXiv.org

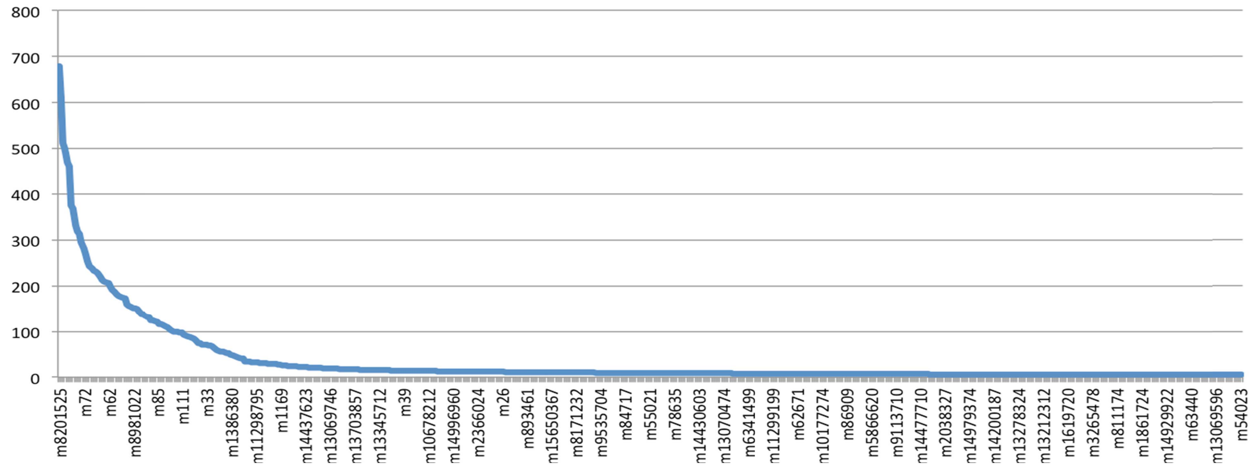

Figure 4. Distribution of entities in the collection.

the top 10 topics and from each topic the top 5 most probable mathematical expressions to select the final set of entities for the second step of our model. In all the experiments the Gibbs sampler was run 1,000 iterations. In Figure 4 we show the distribution of the frequencies of the top 3000 entities. This figure is similar to a word distribution over corpora and motivates the use of tf-idf for the selection of the vocabulary. After pruning the entities found in the collection we proceeded to use LDA over the entities.

4.3 Semantic representation

Once we have the set of relevant objects our next task is to find the semantics of these objects by using another probabilistic model. In our approach outlined in section 3.2 we mentioned that for finding the semantics given the set of objects to use, we need to account for two parameters: the size of the context window and the number of senses that each object can have in the collection.

We argue that to find satisfactory values of these parameters we need to advocate for an iterative application and data driven approach. Therefore, following successful results from the natural language community for nouns and verbs inductions, we run several models and choose the one that best fits our task. Indeed, the selection of these two parameters will vary depending on the task. In our case study, we want to see if we can help to predict the primary MSC class for our collection. In the following section we describe the models used for the prediction task. Since our interest is to investigate the effect of the context window size and the number of senses, we fixed the parameters of the LDA model $\boldsymbol{\alpha}$ to .01, β to 0.1, and the number of iterations for the Gibbs sampler was set to 2,000 in all the experiments.

4.4 Evaluation and results

We generated 12 models with different combinations of the two parameters of our interest: the number of senses and the context size window. For the number of senses we use: 3, 5 and 7. And for each one we generate four window size corpora: 7, 9, 13 and 17 from the collection $\mathbf{C}_{\text{trans}}$. With this setting and given the number of entities selected from Section 3.1 –500, we proceed to compute LDA as mentioned in step 2 section 3.2. And after that we carry on to the task of predicting the top MSC class of our original collection $\mathbf{C}_{\text{entities}}$.

We used Support Vector Machines [20] as our machine learning algorithm and applied ten-fold cross-validation for model selection. Results for this classification task are shown for each sense-context combination in Tables 1, 2 and 3 and in Figure 5.

Table 1. F1 scores for 3 senses

Context size	F1 Score
Window length of 7 words	.3307
Window length of 9 words	.3180
Window length of 13 words	.3203
Window length of 17 words	.3320

The results might look displeasing since ideally one would like to have nearly human performance in this type of classifications task. Information providers are eager when solutions do seem to accomplish nearly perfect results automatically. However, if we abstract for a moment and reflect about the results, lets remember the fact that this is just one entity –relevant from a document, a scientific document, in which the language is far more difficult to model than simpler scenarios such as a news. And yet another important matter to keep in mind is that people perform this task today –experts trained for the task. Therefore, to interpret the results we need to inspect how the model captures the semantics of the mathematical expressions. We show five mathematical monomial expressions with their most likely keywords in tables 4, 5, 6 and 7. For each case we selected the dominant sense of each context size window from the probability distribution of the topic model.

Table 2. F1 scores for 5 senses

Context size	F1 Score
Window length of 7 words	.2867
Window length of 9 words	.2730
Window length of 13 words	.4440
Window length of 17 words	.2740

Table 3. F1 scores for 7 senses

Context size	F1 Score
Window length of 7 words	.2443
Window length of 9 words	.2500
Window length of 13 words	.2483
Window length of 17 words	.2543

The model has been able to capture the use of the mathematical expressions even though as features to predict the taxonomy class of a document they have failed. Interesting to notice is that most of the entities selected by the probabilistic model are short mathematical expressions. And in a way it makes sense since these are the building blocks for more complex but yet very specific mathematical equations very often described by these simple expressions. It is remarkable that even this intuition has been been assimilated by the probabilistic model. Now, to continue in our journey we tackle two questions: first, how good are the semantics encoded in the mathematical expressions in this particular classification task? should we continue and generate more models?

Table 4. Keywords for mathematical expression k

Context sense	Top keywords
Three senses	function, theorem, model
Five senses	algebra, function, equation
Seven senses	function, algebra, equation

Table 5. Keywords for mathematical expression A

Context sense	Top keywords
Three senses	equation, set, lemma
Five senses	theorem, set, lemma
Seven senses	nahm, assume, lemma

Table 6. Keywords for mathematical expression g

Context sense	Top keywords
Three senses	ideal, model, algebra
Five senses	space, function, graded
Seven senses	prime, finite, model

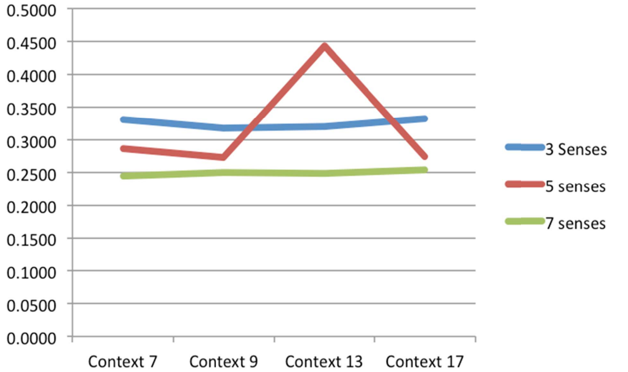

Figure 5. Summary of SVM Classification.

To investigate these issues, we need to depend on statistics analysis. Therefore, the question we need to answer is the following: how good are these semantics to help in the classification process? To gain an intuition about it, we implemented another model based on the title and abstract of each document in the collection as a baseline. We implemented a simple model, so no deep learning for words or phrases in this first approximation. Thus, we use the "bag of words" assumption as our document model; and formulae found in abstracts were ignored –treated as stop-words. Moreover we joined the text of the title and the abstract and therefore the model knows nothing about these two metadata elements, only its content as one single element: plain bag of words. We then perform some basic preprocessing: lower case conversion, stemming and omitting terms below length four to prune the vocabulary size, and terms that occur less than five times in the corpus. Finally, we use the tf-idf scores to further select our final vocabulary size and proceed with the classification task. Again, we use SVM with ten-fold crossvalidation for model selection. The results on the classification task were .42. This result is slightly less than our best model. To conclude whether this difference is significative we performed a Wilcoxon signed rank test over the F1 scores per class between the two models. With a p-value of 1.597e-05 we conclude that statistically speaking the difference is significative. These findings are encouraging. We can argue about at least two conclusions from our classification task: one, mathematical expressions can help in a classification task such as predicting the MSC classes from a document collection. Two, even though they are relevant as entities, in order to improve in the task one can imagine applying our model but looking at another type of entities, perhaps theorems. In fact, the results help to understand the role of mathematical expressions for this particular task. We can argue that they are not the first class citizens but they do represent valuable assets. To further determine what other entities are worth modeling to solve more accurate this particular task, one can gain insight probably through a user study. This user study can be helpful to discover what other entities are relevant in this domain and use our model to investigate if the findings for such a user study correlates to what the model can perform for this particular task.

5. CONCLUSIONS AND FUTURE WORK

In recent years the efforts to provide quality of metadata in the growing number of scientific digital library collections has heightened the need to exploit first class citizens: entities. Indeed, work has increased in the area of entity recognition. And the idea is to find both through manual work or automatic procedures different types of entities. Today we can find evidence in domains such as in chemistry, biomedicine, physics, and mathematics and definitely seeking is entity centric and it is here to persist. But when one makes a pause to think about this issue we can imagine a point in which we ask ourselves: what are these entities? How can we represent them? How can we find them? And how can use them? Surely, one central piece of work missing in the literature is the analysis of the semantics of these entities. And our work presented in this paper has provided new insights to bridge this profound gap.

The idea behind of our solution is to understand how to represent the semantics of these entities and –equally important how to find them in a given collection. Since there is uncertainty in both tasks we have chosen a probabilistic data driven approach. We use a probabilistic topic model that even though is simple, has positively helped us to bring valuable findings to the community. Furthermore, our findings confirm that data driven approaches should be considered as a valid framework to deal with the challenge of making sense of current trends –such as entity search to infer and represent this explosion of knowledge generation.

Nevertheless there is still some room for improvement. In particular, for the first step of our model –finding relevant objects, some empirical tests need to be performed to account for cases in which there is no knowledge to use when trying to initiate the model with a plausible number of topics. Even though experimenting with this parameter can yield to a solution, using a nonparametric Bayesian method could be time saving and perhaps obtain a better solution. For instance, one can use [21]. Therefore, in the near future we will investigate the effect of using such methods.

Another direction of work relates to the selection of the number of senses for each entity. In this work we have used the same for all the entities of interest, however, further work remains to be done to find for each entity a number of senses that yields an optimal solution. We anticipate to work on a solution for this problem that can be general enough to consider a data driven approach, perhaps with some sort of network analysis and/or user feedback provided by query logs.

Finally, another course we would like to explore is to combine more than one observed entity in the probabilistic model to investigate if such a model complexity can provide us with better results for a particular application.

6. ACKNOWLEDGMENTS

The German Research Foundation supported this work (DFG) within the Qualimeta Project.

7. REFERENCES

[1] D. M. Blei, A. Y. NG, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research. 2003

[2] Blei, D. M., & Lafferty, J. D. (2009). Topic Models. In Text Mining: Classification, Clustering, and Applications (pp. 71–89). Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. doi:10.1145/1143844.1143859

[3] Blei, D. M. (2012). Introduction to Probabilistic Topic Modeling. Communications of the ACM, 55, 77–84. doi:10.1145/2133806.2133826.

[4] Goulart, R. R. V., Strube de Lima, V. L., & Xavier, C. C. (2011). A systematic review of named entity recognition in biomedical texts. Journal of the Brazilian Computer Society. doi:10.1007/s13173-011-0031-9.

[5] Settles, B. (2005). ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics, 21, 3191–3192. doi:10.1093/bioinformatics/bti475

[6] Filippov, I. V., & Nicklaus, M. C. (2009). Optical structure recognition software to recover chemical information: OSRA, an open source solution. Journal of Chemical Information and Modeling, 49, 740–743. doi:10.1021/ci800067r

[7] Lowe, D. M., Corbett, P. T., Murray-Rust, P., & Glen, R. C. 2011. Journal of Chemical Information and Modeling, 51, 739–753. doi:10.1021/ci100384d

[8] Park, J., Rosania, G. R., Shedden, K. A., Nguyen, M., Lyu, N., & Saitou, K. (2009). Automated extraction of chemical structure information from digital raster images. Chemistry Central Journal, 3, 4. doi:10.1186/1752-153X-3-4

[9] P. Sojka and M. Lška. The Art of Mathematics Retrieval. Proceedings of the ACM Conference on Document Engineering. 2011

[10] Michael Kohlhase, Bogdan A. Matican, and Corneliu C. Prodescu. MathWebSearch 0.5 -Scaling an open Formula Sarch Engine. Conferences on Intelligent Computer Mathematics (CICM). 2012

[11] Kamali, S., & Tompa, F. W. (2013). Retrieving documents with mathematicalcontent. In Proceedings of the 36th international ACM SIGIRconference on Research and development in information retrieval – SIGIR '13 (p. 353). doi:10.1145/2484028.2484083

[12] Sun, B., Mitra, P., & Giles, C. L. (2008). Mining, indexing, and searching for textual chemical molecule information on the web. In Proceeding of the international conference on World Wide Web (pp. 735–744). doi:10.1145/1367497.1367597

[13] Tönnies, S., Köhncke, B., Koepler, O., & Balke, W.-T. (2010). Exposing the Hidden Web for Chemical Digital Libraries. In Int.l Joint Conference on Digital Libraries (pp. 234–244). doi:10.1145/1816123.1816159

[14] Vickrey, D., Biewald, L., Teyssier, M., & Koller, D. (2005). Word-Sense Disambiguation for Machine Translation. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05) (pp. 771–778). doi:10.3115/1220575.1220672

[15] Carpuat, M., & Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 61–72. Retrieved from papers2://publication/uuid/CA8E0BC3-96B6-4123-8674-4E4BD98AACA9

[16] Brody, S., & Lapata, M. (2009). Bayesian Word Sense Induction. Computational Linguistics, 103–111. doi:10.3115/1609067.1609078

[17] Lau, J. H., Cook, P., McCarthy, D., Newman, D., Baldwin, T., & Computing, L. (2012). Word sense induction for novel sense detection. In Proceedings of the 13th Conference of the European Chapter of the Association for computational Linguistics (EACL 2012) (pp. 591–601).

[18] Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis (special Volume of the Philological Society), 1952-59, 1–32.

[19] Griffith TL, Steyvers M (2004). Finding Scientic Topics. Proceedings of the National Academy of Sciences of the United States of America, 101, 5228-5235

[20] Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Knowledge Discovery and Data Mining, 2, 121–167. Retrieved from /papers/Burges98.ps.gz

[21] Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. Journal of the American Statistical Association. doi:10.1198/016214506000000302