# Thesis Proposal

Alexey Grigorev

January 7, 2015

| | |
|---|---|
| Official Affiliation: | IT4BI Master Thesis @ TU Berlin / DIMA |
| Title of the thesis: | Identifier namespaces in mathematical notation |
| Candidate: | Grigorev, Alexey |
| Matriculation number: | 0367628 |
| Advisor: | Schubotz, Moritz @ TU Berlin |
| Co-advisor: | Soto, Juan @ TU Berlin |
| Official Supervisor: | Prof. Dr. Markl, Volker @ TU Berlin |
| Planned period: | February 2015 - August 2015 |
| Version: | 0.8 |

# 1 Background

## 1.1 Introduction

In computer science, a *namespace* refers to a collection of terms that are managed together because they share functionality or purpose, typically for providing modularity and resolving name conflicts [1]. For example, XML uses namespaces to prefix element names to ensure uniqueness and remove ambiguity between them [12], and the Java programming language uses packages to organize identifiers into namespaces for modularity [14].

In this thesis we will extend the notion of namespaces to mathematical formulae.

In logic, a *formula* is defined recursively, and, in essence, it is a collection of variables, functions and other formulas, and formally the symbols for the variables and functions can be chosen arbitrarily [16]. However, in contrast to first order logic, in this work we are interested in the symbols in formulae and in mathematical notations that are used by different research communities. For example, in physics it is common to write the energy-mass relation as $E = mc^2$ rather than $x = yz^2$. However, the same identifier may be used in different areas but denote different things: For example, $E$ may refer to "energy", "expected value" or "elimination matrix", depending on the domain of the article. Thus, we can note that these identifiers form namespaces, and we refer to such namespaces as *identifier namespaces*, and to the process of discovering identifier namespaces as *namespace disambiguation*.

In this thesis we compare different approaches for namespace disambiguation. The first approach is to assume that there is a strong correlation between identifiers in a document and the namespace of the document, and this correlation can be exploited to categorize documents and thus discover namespaces. For example, if we observe a document with two identifiers $E$, assigned to "energy", and $m$, assigned to "mass", then it is more likely that the document belongs to the "physics" namespace rather than to "statistics". To use it, we need to map identifiers to their definitions, and this can be done by extracting the definitions from the text that surrounds the formula [2]. Other approaches are based on the text of the documents, rather on the formulae [10], but nonetheless we believe that there is a correlation between the textual content of the document and the namespace of its identifiers.

### 1.2 Related Work

Kristianto et al [3] highlight the importance of interlinking the scientific documents and in their study they do it through annotating mathematical formulae and finding the documents that share the same identifiers. Schöneberg et al [9] propose mathematical-aware part of speech tagger and they discuss how it can be applied for classifying scientific publications.

There are several researches related to extracting textual description of mathematical formulae. One of the earliest works is by Grigore et al [4] that focuses on disambiguation, and Yokoi et al [5] that focuses on advanced mathematical search.

Pagel and Schubotz [2] suggest a Mathematical Language Processing framework - a statistical approach for relating identifiers to definitions. Similar approach is suggested in [5], [3] and [8], where the authors use machine learning methods for extracting the definitions.

Some work is also done in clustering mathematical formulae by Ma et al [7] to facilitate formula search where they propose features that can be extracted from the formulae.

In computational linguistics there is a related concept called *semantic field* or *semantic domain*: it describes a group of terms that are highly related and often are used together. Words that appear frequently in same documents are likely to be in the same semantic field, and this idea is successfully used for text categorization and word disambiguation [11].

## 2 Goals

The main objective of this study is to discover identifier namespace in mathematical formulae. We aim to find *meaningful* namespaces, in the sense that they can be related to a real-world area of knowledge, such as physics, linear algebra or statistics.

Once such namespaces are found, they can give good categorization of scientific documents based on formulae and notation used in them.

We believe that this may facilitate better user experience: for instance, it will allow users to navigate easily between documents of the same category and see in which other documents a particular identifier is used, how it is used, how it is derived, etc. Additionally, it may give a way to avoid ambiguity. If we follow the XML approach [12] and prepend namespace to the identifier, e.g. "physics.$E$", then it will give additional context and make it clear that "physics.$E$" means "energy" rather than "expected value".

We also expect that using namespaces is beneficial for relating identifiers to definitions. Thus, as an application of namespaces, we would like to be able to use them for better definition extraction. It may help to overcome some of the current problems in this area, for example, the problem of *dangling identifiers* [2] - identifiers that are used in formulae but never defined in the document. Such identifiers may be defined in other documents that share the same namespace, and thus we can take the definition from the namespace and assign it to the dangling identifier.

To achieve these objectives we define the following research tasks:

1. To identify similarities with computational linguistics, computer science and mathematics
2. To study existing solutions for clustering textual and mathematical data and how to use them to discover meaningful namespaces
3. To implement promising approaches to namespace disambiguition
4. To evaluate these approaches in order to find the best
5. To incorporate the found namespaces to the existing MLP framework (described in [2])

These tasks are explained in details in the next section.

# 3 Realization

## 3.1 Namespace disambiguation

To accomplish the proposed goal, we plan the following.

First, we would like to study and analyze existing approaches and recognize similarities and differences with identifier namespaces. From the linguistics point of view, the theory of semantic fields [15] and semantic domains [11] are the most relevant areas. Then, namespaces are well studied in computer science, e.g. in programming languages such as Java [14] or markup languages such as XML [12]. XML is an especially interesting in this respect, because it serves as the foundation for knowledge representation languages like OWL (Web Ontology Language) [13] that use the notion of namespaces as well.

The process of manual categorization of mathematical corpus is quite time consuming. What is more, scientific fields are becoming more and more interconnected, and sometimes it is hard even for human experts to categorize an article. Therefore, we believe that the namespaces should be discovered in an unsupervised manner.

Thus, we would like to try the following methods for finding namespaces: categorization based on the textual data [10], on semantic domains [11], on keywords extracted from the documents [9] or on definitions extracted from the formulae in the documents [2].

The data set that we plan to use is a subset of English wikipedia articles - all those that contain the `<math>` tag. The textual dataset can potentially be quite big: for example, the English wikipedia contains 4.5 million articles, and many thousands of them contain mathematical formulae. This is why it is important to think of ways to parallelize it, and therefore the algorithms will be implemented in Apache Flink [6].

## 3.2 Evaluation

The meaningfulness of discovered namespaces can be evaluated by sampling some documents of the same category and examining them manually. We expect that within one discovered category we should be able to observe documents that can be related to some real-world domain. For example, if we sample two documents and get "Ordinary Least Squares" and "Kernel Regression", then we can relate them to the same area (e.g. "Statistics") and this is the result we would like to achieve. On the other hand, if we observe "Ordinary Least Squares" and "Dirac comb" within the same category, then it will make it harder to explain such categorization.

Many wikipedia articles have been manually categorized and it is possible to exploit that. For example, many articles on Machine Learning contain a special macro {{`Machine learning bar`}} that renders a list of links to related articles. Thus, if we see such a macro at one page, we expect to observe it on another page within the same namespace. Unfortunately, the way of categorizing is not always consistent, and in some cases the macros look quite differently. For example, for statistical articles the macro is {{`Statistics|correlation|state=collapsed`}}. This makes it impossible to use it for automatic evaluation, however, it does provide good help in manual evaluation of results.

Additionally, we plan to see to what extent the namespaces are beneficial for the keyword extraction, and therefore, we plan to incorporate them into the MLP framework [2] to see if the results give better precision and recall. Thus, the results by Pagel and Schubotz [2] will serve as the baseline for this evaluation.

Lastly, it can also be interesting to take advantage of so-called *interlanguage links* that link one page in wikipedia in one language to the equivalent pages in another languages. Mathematical notation may be consistent across multiple languages, and this can be used in the evaluation of the results. For example, we can take an article and check how similar are discovered namespaces in different wikipedias. Furthermore, it is also possible to make use of machine

translation techniques and see whether the description of common identifiers are the same or not.

# 4 Bibliography

# References


[1] Duval, E., Hodgins, W., Sutton, S., Weibel, S. L. (2002). Metadata principles and practicalities. *D-lib Magazine*, 8(4), 16.

[2] Pagael, R., Schubotz, M. (2014). Mathematical Language Processing Project. *arXiv preprint* arXiv:1407.0167.

[3] Kristianto, G. Y., Aizawa, A. (2014). Extracting Textual Descriptions of Mathematical Expressions in Scientific Papers. *D-Lib Magazine*, 20(11), 9.

[4] Grigore, M., Wolska, M., Kohlhase, M. (2009). Towards context-based disambiguation of mathematical expressions. In *The Joint Conference of ASCM* (pp. 262-271).

[5] Yokoi, K., Nghiem, M. Q., Matsubayashi, Y., Aizawa, A. (2011). Contextual analyis of mathematical expressions for advanced mathematical search. *Polibits*, (43), 81-86.

[6] Apache Flink, http://flink.incubator.apache.org/

[7] Ma, K., Hui, S. C., Chang, K. (2010). Feature extraction and clustering-based retrieval for mathematical formulas. In *Software Engineering and Data Mining* (SEDM), 2010 2nd International Conference on (pp. 372-377). IEEE.

[8] Kristianto, G. Y., Nghiem, M. Q., Matsubayashi, Y., Aizawa, A. (2012). Extracting definitions of mathematical expressions in scientific papers. In *The 26th Annual Conference of JSAI*.

[9] Schöneberg, U., Sperber, W. (2014). POS Tagging and its Applications for Mathematics. In *Intelligent Computer Mathematics* (pp. 213-223). Springer International Publishing.

[10] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys* (CSUR), 34(1), 1-47.

[11] Gliozzo, A., Strapparava, C. (2009). Semantic domains in computational linguistics. Springer.

[12] Bray, T., Hollander, D., Layman, A. (1999). Namespaces in XML. *World Wide Web Consortium Recommendation REC-xml-names-19990114*. http://www.w3.org/TR/1999/REC-xml-names-19990114.

[13] McGuinness, D. L., Van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation*, 10(10), 2004.

[14] Gosling J., Joy B., Steele G., Bracha G., Buckley A. (2014) The Java Language Specification, Java SE 8 Edition. In *Java Series*. Addison-Wesley Professional.

[15] Vassilyev, L. M. (1974). The theory of semantic fields: A survey. *Linguistics*, 12(137), 79-94.

[16] Barwise, J., Etchemendy, J., Allwein, G., Barker-Plummer, D., Liu, A. (2000). Language, proof and logic. CSLI publications.

City, Date, Signature of the student

City, Date, Signature(s) of the advisor(s)