Semantic Search for Quantity Expressions
Guided Research Proposal

Tom Wiesing
Supervisor: Michael Kohlhase
Jacobs University, Bremen, Germany

March 1, 2015

Abstract

In this proposal we describe how to introduce units to MathWebSearch. The aim of the project is build a complete semantics-aware system that searches a corpus of documents for quantity expressions. The project will be based on the existing MathWebSearch system and related technologies. [1]

EdN:1

1 Introduction

In this paper we want to give an approach to Semantic Search for Quantity Expressions. A quantity expression is a value together with a physical unit, for example $25\frac{\text{m}}{\text{s}}$ where 25 is the value and $\frac{\text{m}}{\text{s}}$ the unit. This quantity expression is equivalent to $90\frac{\text{km}}{\text{h}}$ (with equivalent in this sense meaning it expresses the same quantity) although the units are different. In a semantic search for quantity expressions we want to be able to search for a certain quantity expression and find any equivalent ones as well.

MathWebSearch (MWS for short) is an existing semantics-aware system to search LaTeX documents [1] for mathematical formulae [HKP14]. As it is semantics-aware it not only searches for formulae in a simple-minded "text search" way but also includes simple transformation rules, such as $a + b = b + a$. Additionally it can deal with wildcards such as $x + \sqrt{x}$. For this query MWS would deliver forumlae as above where x has been substituted with any sub-formula.

So far the transformation system has been used by MWS exclusively for mathmatical forumlae. In this paper we propose an extension for quantity expressions which should prove useful for physicists. The end user will search, for example, $100\,^{\circ}\text{C}$ and also get results which show 212°F or 373.15K.

This proposal is organised as follows: In section 2 we shortly describe the existing MathWebSearch system and then proceed in section 3 to describe in detail the proposed extension. Finally in section 4 we discuss challenges as well as related work.

---

[1] EdNote: Spin

[1] Technically, MWS itself can only search XHTML documents. However with the help of LaTeXml [Mil] it also handles LaTeX documents.

2 An Overview of the existing MathWebSearch system

2.1 The Components of MathWebSearch

As mentioned above, MathWebSearch is a search engine for forumlae. MWS consists of 3 main components as well as a frontend [2] [KP].

The backend consists of 3 main components,

1. a crawler,

2. a core system and

3. a public rest API.

The crawler, as its name suggests, crawls corpera for forumlars. For each corpus MWS uses, a seperate crawler has to be implemented. The crawled formulae are passed to the core system and indexed in an MWS Harvest. The core system is also responsible for parsing queries and sending results back to the REST API. This is done by searching the harvest only.

Because MWS is semantics-aware, the harvest can not only contain the exact formulars that are found in the original corpus but also all versions that are quivalent to it. These are generated with the help of MMT and theory graphs.

The frontend for MathWebSearch, which is not part of MWS itself but running client-side in a web browser, is written in HTML5, CSS and JavaScript. It accesses the REST backend and depends on MathML support to render Mathematics. When the client enters a LATEX formular to search for, the LATEXml daemon [Gin] is used to transform the query in content MathML. Next, the client renders the MathML (to show the formular the user is searching for) and then sends the query off to the MWS API. Upon receiving results, the client renders them and links to the original documents.

There are several implementations of frontends and crawlers as well as extensions of Math-WebSearch. One particular implementation is capabale of crawling the arXMLiv corpus, which contains approximatly 750.000 documents. A list of demos can be found at [Mat].

2.2 Modeling of Mathematical Theories with MMT

MMT is a **M**odule system for **M**athematical **T**heories [RK13]. In MMT we can represent mathematical theories, views and imports.

In MMT, a mathematical theory in this sense is a set of symbols along with a set of axiomes[3] [2]. In their MMT representation both the symbols and axioms have types.   EdN:2

Theories can be related in two ways, with imports and views. If theory A imports theory B then theory A contains all symbols and axiomes from theory B. As an example, the theory of a *commutative groups* imports the theory of *groups*. The *commutative group* theory additionally has the axiom of commutativity. We can consider an import as a type of inheritance relation. In contrast to this a view is a type of interpretation relation. For example, there is a view from the theory of *Monoids* to the theory of *Natural numbers*. A theorem in one theory can

---

[2] The frontend is not part of MWS directly but rather built on top of the REST API, more on this later.

[3] While in mathematics axiomes do not require proof, in MMT it is possible for an axiom to actually be a theorem that has a proof.

[2] EDNOTE: Explain this MUCH MUCH BETTER, this way it is wrong.

be translated along views as well as along imports. For example, given a theorem in the theory of groups, we automatically get a theorem in the theory of commutative groups.

2.3 Using Theory Graphs

The theories and their relations can be represented with the help of a graph. This graph is then called theory graph. In figure 1 we can see a simple example of a theory graph.
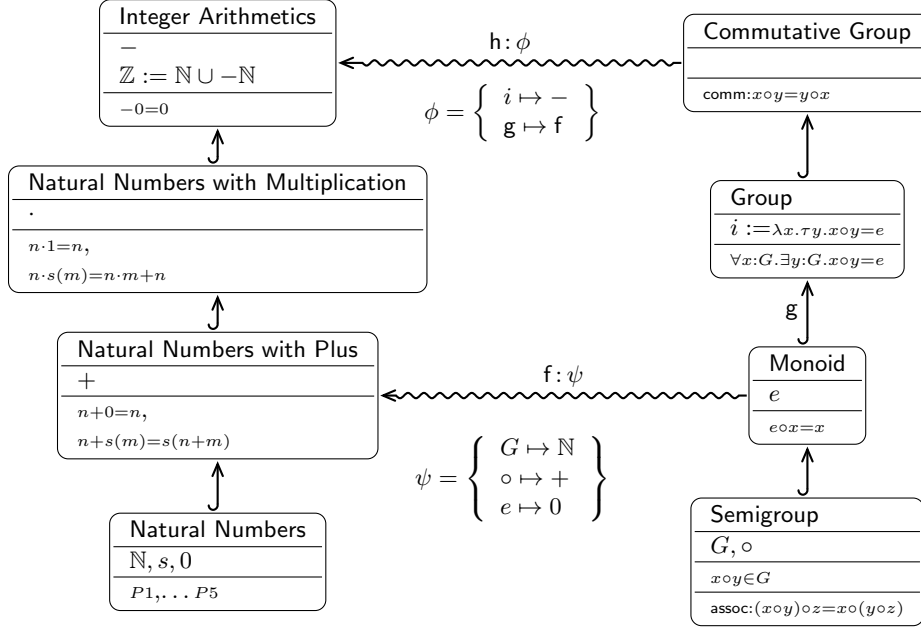

Figure 1: A simple theory graph. Imports are represented as solid edges and views as dashed edges.

Every theory in MMT has a (globally unique) URL, called the MMT URL. Sometimes it is useful to not write down a theory explicitly, but only give an existing *represented theory* and a view to a new theory [3]. We then call this new theory an *induced theory*. The MMT URL of induced theories is given via the view only and not represented explicitly. However, the MMT URL contains enough information that MMT can generate all symbols, axioms and induced theorems explictly [IKP14] [4].   EdN:4

In an MWS Harvest we exploit this property of induced theorems. We have several represented theories, the forumlae in the corpus, and many more induced theories, other equivalent representations of these formulae. With the help of MMT, the MMT Harvest is built by representing all theses theories (forumlae) explictly.

EdN:3

---

[3]EDNOTE: This is theorems only
[4]EDNOTE: Mention URKs

3 The goal - A Semantic Search For Quantity Expressions Based On MathWebSearch

The goal of the guided research is to get a complete system that can perform a semantic search of quanity expressions on a specific corpus. As such it needs 4 components: (1) *A crawler* that searches through[5] documents and finds units which can be indexed, (2) *a core system* that takes care of storing indexes and processing queries and is aware of units with the help of a theory graph, (3) *a rest API* that has to have a way of sending units from and to a client and (4) *a frontend* that runs in the web browser.   EdN:5

Furthermore the system will be flexible with respect to

- *the quanity expressions it understands,* Adding new units should be a simple process so that the system can be easily extended to also capture rare units and

- *the corpus* which should be exchangable easily.

The crawler will be based on existing MWS crawlers. As it is very difficult to find and mark up all units in a corpus automatically, we will restrict ourself to an artificial corpus containing only lorem ipsum [6]documents with randomly inserted units. This does not solve   EdN:6 the problem of writing a crawler, however it will enable us to build a proof-of-concept system first.

The core system will almost completly be inherited from existing implementations. As it does not search formulae, it has to be made aware of unit translations. This will be done with the help of theory graphs. It is not yet clear how exactly these will look, a speculation can be found below in 4.

The rest-based API will have to take the user input from the client and pass the entire quantity expression on to the core system. Once it receives a result, it will have to translate these back into human-readable form and then pass these on to the frontend.

As with the existing system, the frontend will be a web page that incorperates 3 main elements, (1) a search input for a value (2) a search input for a unit and (3) a result page that displays results and links to the found documents.

4 Challenges and related Work

A bigger challenge with the approach will be to find a standardised representation for quantity expressions, mainly for units[4] [7].   EdN:7

Because we want to translate between these representations however, it is insuffcent to just have such a representation. We additionally need to find a standardised way of translating between units need to exist as well. Because MWS uses Theory Graphs and MMT to translate between equivalent formulae, we will have to write down all units in the form of a theory graph. Translations between units will be represented as views.

To translate between units we could use simple translation formulae such has $x\text{K} = x + 273.15°C$. However because we want to support composite units (such as meters per second

---

[5]EDNOTE: reformulate this

[6]EDNOTE: What is my contribution, the exact details are not important.

[4]Here, standardised means that all quantity expressions are represented in a standardised, machine-readable fashion, not equivalent quantity expressions being representeted in exactly one way.

[7]EDNOTE: MathML note OM CDs

$\frac{m}{s}$) as well, this can cause problems. In machine representation [8], every quantity expression has to have a finite precision representation. When translating between 2 composite units, this can easily cause rounding errors. In particular, when an author gives an approximation of a quantity in a document, they might round differently depending on the units used. EdN:8

Hence it will not be sufficient to use exact translations. Thus we should search for a range of values instead. This has recently been implemented as an extension to MathWebSearch [Ham]. It remains to be seen how exactly these ranges should be used.

Furthermore the units will have to be entered in some fashion [9] in the frontend. While it is trivial to design an interface where a single unit can be entered, it is non-trivial when we want to recognise composite units as well. Furthermore units with (si-)prefixes (such as kilo or mega) can either be recognised seperatly or as part of the unit (multipling directly into the value of the quantity expression). There are several input formats that can be used: AsciiMath, LaTeX and MathML, to name only a few. Independent of the input format, the end result (delivered to the backend of the search engine) will have to be represented in one way. EdN:9

Finally, there is the problem of choosing a suitable corpus and implementing a crawler which will find and mark up quantity expressions. The latter will be taken on by Stiv Sherko in a seperate effort [She14]. The task of finding a suitable corpus is postponed for now. In case there is sufficient time a ycorpus can be marked up manually and crawled afterwards.

References

[Gin]   Deyan Ginev. TheLaTeXml daemon: Editable math for the collaborative web.

[Ham]   Radu Hambasan. Ranges in mathwebsearch. private communication.

[HKP14] Radu Hambasan, Michael Kohlhase, and Corneliu Prodescu. MathWebSearch at NTCIR-11. In Noriko Kando and Hideo Joho andKazuaki Kishida, editors, *NTCIR 11 Conference*, pages 114–119, Tokyo, Japan, 2014. NII, Tokyo.

[IKP14] Mihnea Iancu, Michael Kohlhase, and Corneliu-Claudiu Prodescu. Representing, archiving, and searching the space of mathematical knowledge. In Hoon Hong and Chee Yap, editors, *Mathematical Software - ICMS 2014 - 4th International Congress*, volume 8592 of *Lecture Notes in Computer Science*, pages 26–30. Springer, 2014.

[KP]   Michael Kohlhase and Corneliu Prodescu. MathWebSearch manual. Web manual, Jacobs University.

[Mat]   Math WebSearch a semantic search engine. web page at http://search.mathweb.org. seen September 2008.

[Mil]   Bruce Miller. LaTeXML: A LaTeX to XML converter.

[RK13]   Florian Rabe and Michael Kohlhase. A scalable module system. *Information & Computation*, 0(230):1–54, 2013.

---

[8]EDNOTE: Seperate problem
[9]EDNOTE: The same?

[She14]   Stiv Sherko. Extracting quantity and unit semantics from technical documents. Bachelor thesis proposal, Computer Science, Jacobs University, Bremen, 2014.