$5e^{x+y}$: A Math Aware Search Engine (for CDS)

Arthur Oviedo
EPFL, Switzerland.
arthur.oviedo@alumini.epfl.ch

Nikolaos Kasioumis
CERN, Switzerland.
nikos.kasioumis@cern.ch

Karl Aberer
EPFL, Switzerland.
karl.aberer@epfl.ch

ABSTRACT

This paper presents $5e^{x+y}$, a system that complements CERN Document Server (CDS), by adding extracting, indexing and querying of mathematical content, expressed as mathematical equations.

Categories and Subject Descriptors

X.X [**Digital Libraries**]: Miscellaneous; Y.Y [**Information Retrieval**]: Metrics

General Terms

Theory

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

CDS[3] (CERN Document Server) is the institutional digital library system developed and used at CERN[1]. It contains more than 1 million records and stores more than 400000 full text documents, organized in different collections among published articles, books, preprints. Invenio[4] is the open-source digital library software platform behind CDS and is developed in parallel to CDS at CERN. Besides CDS, Invenio supports around thirty scientific institutions worldwide including INSPIRE (a collaboration between Fermilab, CERN, DESY and SLAC) EPFL and ILO. An important part of the CDS content contains mathematical content, represented in the form of equations. Only the Preprints collection contains more than 700000 records, harvesting documents from services like ArXiV[2] where most of the documents are in the areas of physics, mathematics and statistics which are very rich in this type of content. Even though Invenio provides a powerful searching mechanism that allows to query the stored records by different fields like author or keywords, these approaches are not suitable to query this type of content. $5e^{x+y}$ is developed as a first attempt to address this limitation by allowing querying for records based on mathematical equations.

JCDL 2014 Knoxville, USA

2. SYSTEM OVERVIEW

In this section we provide a general overview of the main design points of $5e^{x+y}$. A more detailed description of the system can be found in [7].

2.1 Features Extraction

Our initial focus in the project was to define a good set of features that we could extract and index from our equations. We divided them, into two different categories. The notational features relate to the actual naming and representation of the different elements of the equation. The selected format for representing and internal processing of the equations was MathML because it allows an easier processing.

For notational features, we extracted the leave elements of the equation: variables, numbers and operator represented by the tags <mi>, <mn> and <mo> respectively. We indexed as well 2-grams and 3-grams of symbol-operator and symbol-operator-symbol. We perform some processing of these tokens such as normalizing and decomposing of complex characters. This allowed us to partially match characters with different unicode points but similar or equal visual rendering. For instance LATIN CAPITAL LETTER A WITH RING ABOVE with code 0xc5 and ANGSTROM SIGN with code 0x212b both are rendered as Å and both can be used to refer to the Angstrom measurement unit. Following the example, we indexed both the original token <mi>Å</mi> and the normalized form <mi>Å</mi> <mi>A</mi>. Another set of rules that we applied take into account the different types of operators that are semantically related but whose unicode representation does not provide meaningful ways to match them. We grouped related operators, for example the INTEGRAL character with code 0x222b generates the tokens <mo>$\int$</mo> and INT_OP and the CONTOUR INTEGRAL character with code 0x222e generates the tokens <mo>$\oint$</mo> and INT_OP. In this way, both have a partial matching in the second token.

The second set of features relate to the structure of the equation. Two different properties of mathematical equations impose additional challenges for a MIR system. First, variables can be used indistinguishable and any character set of characters can be used to represent the same equation. $f(t) = vt + c$ and $g(x) = ac + b$ can represent the same linear equation though the context in which each variable is used follows certain conventions. For addressing this issue,

we relied on pattern matching, a very powerful feature that current Computer Algebra System (CAS) provide. We integrated Mathematica into our system, to perform pattern matching on the equations and identify the occurrence of a predefined of patterns from a predefined set. This allows us to detect if the equation contained certain types of common algebraic structures. For instance, the pattern `a_.x^ 2 + b_.x + c_.` is used to identify quadratic expression. This mechanism is very powerful since represent any type of expression, not only a single term.

A second additional challenge that math equations pose is the fact that a given equality can be rewritten in multiple ways taking into account properties of the different operators like associativity, distributivity, commutativity among others. Our approach was to rely again in the features provided by Mathematica and apply the `Simplify` operations which applies a set of predefined set of rules in a equation with the goal of having a unique and concise representation.

2.2 System architecture

$5e^{x+y}$ is implemented in top of the Lucene/Solr framework. We expanded the default schema with several fields, being the most important ones: `math_notational_field` which stores all the described tokens and n-grams, `math_structural_field` which stores the occurrences of predefined algebraic structures and `filename` which is used to store the file/record where the particular equation was found. In order to extend Lucene with the new use case, our work extends some of the base classes as follows: `MultiplePatternTokenizer` is in charge of producing a stream of tokens from a single equations. This tokens are identified by applying multiple regular expressions. `UnicodeNormalizingFilter` applies the unicode normalizing transformation described above. `SynonymExpandingFilter` expands operators, with a manually computed table, with the category they belong to. Finally `StructuralPatternsTokenizer` provides the pattern matching functionality by establishing a communication with Mathematica's `Kernel Link` API.

3. DEMONSTRATION

As part of the demonstration of $5e^{x+y}$, we present a demo instance of the Invenio software with the searching by mathematical equations functionality. The user access the math searching functionality through a link in the main page. Once there, the user is presented with a classic search functionality where the user inputs his equation. He can choose between writing the query on LaTeXor MathML. For writing the LaTeXcode, we include an online editor[5] that renders the the input as the user types it. The first format is normally more familiar to users , however, the internal representation is in MathML since it allows an easier processing. The translation from LaTeXto MathML is done through the third party library SnuggleTex[6] that have some shortcomings, so we allow users to input the query in MathML as well. When the query is ready, the user clicks the submit button and the system performs the search. Once the system has performed the lookup and got the results, these are presented in the results interface. The system shows up the 20 results with highest score. Each result shows the matched equation snippet, the title of the record containing the given equation and the link to the record.

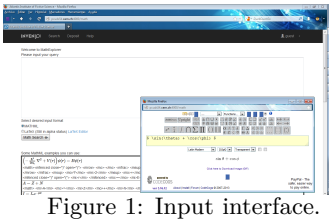

Figure 1: Input interface.

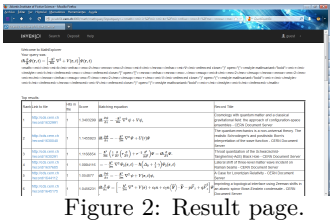

Figure 2: Result page.

4. CONCLUSIONS

5. AVAILABILITY

Both, the Invenio software and the $5e^{x+y}$ module are open source available through GPL Licence. Invenio can be downloaded from `http://invenio-software.org/`. $5e^{x+y}$ can be downloaded from `https://github.com/arthoviedo/cern_math_explorer` .

6. REFERENCES

[1] About cern, 2014. `http://home.web.cern.ch/about`.

[2] arxiv.org e-print archive, 2014. `http://arxiv.org`.

[3] Cern document server, 2014. `http://cds.cern.ch/`.

[4] Invenio, 2014. `http://invenio-software.org/`.

[5] Online latex editor - create, integrate and download, 2014. `http://www.codecogs.com/latex/eqneditor.php`.

[6] T. U. of Edinburgh School of Physics and Astronomy. Snuggletex (1.2.2). `http://www2.ph.ed.ac.uk/snuggletex/documentation/overview-and-features.html`.

[7] A. Oviedo, N. Kaisoumis, and K. Aberer. $5e^{x+y}$: A math-aware search engine for cds. CERN Document Sever.