Languages of Mathematics

Random Walking in the Mathematics of Languages

Petr Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
sojka@fi.muni.cz

Abstract. An essay about mathematics being a sublanguage of other natural languages: how it may be represented, stored, searched and handled in several projects of (European) Digital Mathematics Libraries as DML-CZ or EuDML.
A framework for solving problem of computing of similar papers in a digital library is proposed, allowing several types of similarity type definitions: *plagiarity* counting on common word *n*-grams, *topicality* counting on common topics, or *conarrativity* counting on the same narrative. The vector of the most similar documents for a given similarity type is suggested to be computed using the algorithm by Page for web page ranking, often explained as 'random walking'.

Science is based on trust and integrity. – Venkatraman Ramakrishnan
Nobel laureate 2009

1 Introduction

The language of mathematics can be viewed as a sublanguage of other natural languages. The recent initiatives Towards a Digital Mathematics Library [1,2,3] aim at virtual multilingual digital library with the papers published as peer-reviewed verified archive knowledge in the area of mathematics. The area is well-defined by review databases Mathematical Reviews and Zentralblatt with almost 3,000,000 (metadata and reviews) items of mathematical scientific literature. The integration even on the level of full texts has started, and brings questions like:

- how to represent mathematical language, formulae?
- how to index it, search it?
- how to deal with semantics of mathematics?
- how to classify mathematics, which ontologies to use?
- how to deal with mix of 'informal' texts and formal proofs and specifications?

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 127–133, 2009.

For big software firms like Google (Google Scholar) and ABBYY (FineReader) mathematics is very small niche with very big problems to face. There are, fortunately, several smaller digital mathematics library initiatives like NUMDAM[1] or DML-CZ[2] where new best practices are created and tested, in addition to the development of tools for solving at least some of the problems of handling mathematics. For the final solution of problems like mathematical OCR or semantic representation and searching mathematics handling there are still funds missing, though.

In this paper we shortly sum up current level of understanding of these issues based on the experience of five years of working on the DML-CZ project. We discuss math representation issues in Section 2. We follow with topic of math search and digital libraires in Sections 3 and 4. Finally, in Section 5, we define several kinds of 'similarity' usable not only for mathematical papers and suggest novel framework to compute these general versions of 'similarity' using iterative algorithm used sofar for ranking web pages [4].

> Where possible, the systems will share a common application or database, or perhaps a more common data structure that will allow one system to import / export data with another system without sharing their applications or platforms. – Report by the 511 Interoperability Task Force, April 4, 2005

2 Domain of Mathematics

For communication of mathematics in a digital library, several formats are used: the relations and laws are either expressed verbally in plain language, or formulas and formalisms are used.

On the authoring side, the most widespread and preferred format is the plain TEX's notation or it's markup extensions defined in AMSLATEX. TEX or its successors as pdf(e)TEX are said to be used for the production of more than 90 % of the world's scientific printed journals.

For communication between bots, programs and applications, MathML standard by W3C is supported for mathematics exchange. One can cut and paste formula from Mathematica and paste it in Maple to derive it, and import the result into web page rendered by Firefox.

We can classify the levels mathematics is handled now:

1.0 lexical – words, *strings* of characters or TEX's $ $ notation.
2.0 syntactical – phrases, *parsed* formulas (represented as trees in MathML).
3.0 semantical – *meaning* of parsed phrases (cloud tags/ontologies/OpenMath).

The problem is that the author's message (it's incarnation in the paper's *content* and *form*) does not survive (no standard representation of math) when communicated (via the paper or over the web) to the readers.

Although semantical representations of mathematical formulas in MathML version 3 [3] or in OpenMath's Content Dictionaries [4] are well defined, they are not

[1] http://numdam.org [2] http://dml.cz [3] http://www.w3.org/TR/MathML3/
[4] http://openmath.org

used by authors, probably because there is no strong incentive and benefits for authors. On the opposite, semantical markup gives additional burden to authors to disambiguate their thoughts, when they hurry for publication (Publish or Perish). The cost of semantic-rich markup is not usually willing to be absorbed by publishers – they claim that the price tag is too high. There are estimates that the growth of production costs from standard paper/PDF-only LaTeX to PDF production to LaTeX to validated XML+MathML to LaTeX to PDF is tenfold (from $6 to $60 per page, even if it is outsourced to India or other cheap labour country). Others oppose that it is a must anyway and that by developing authoring tools that take as much of logical markup from author as possible into the source file publisher may leverage the costs to minimum. In the case publisher would not have rich semantically marked XML+MathML+SVG files to build it's services on, it would not be able to compete on the publishing market. Current ability iof some publishers to generate Epub or DAISY formats from their rich XML based representation shows that it actually pays back very quickly. New architectures and services start to appear, based on the rich XML+MathML markup [5].

Quite different requirements have theorem proving systems and computer algebra systems. They use usually their own internal representation of mathematics, with MathML (or LaTeX) as the interface languages.

As simple as possible, but not simpler. – Albert Einstein

3 Search

Neither format mentioned in the previous section is widely accepted and used, though. When one tries to search for citations of Kováčik and Rákosník's paper [6] by Google Scholar,[5] one finds more than a dozen of different citation clusters of it, depending on the OCR errors in this paper author's names and in the 'representation' of math formulas in the paper's title. It may be seen as a clear evidence of current mess of different ways of mathematics representation and treatment. There are attempts to sort out this mess, ambitions of e.g. *Math WebSearch*[6] are much higher.

The widely used *Google Search* only pays attention to the ranking when delivering (math) search results – there is no sign of math representation or disambiguation. *SearchPoint*[7], on the other hand allows walking in the meaning spaces: in the clusters of related pages with different meanings of terms in the question posed.

Mathematical search has both many specifics [7] and many common problems of information retrieval:

- Mathematical notation is context-dependent, e.g. binomial coefficients has different form in different languages and language contexts: $\binom{n}{k}$, ${}_nC^k$, C^n_k, C^k_n all denote the same semantically equivalent notion.
- Identical presentations can stand for multiple distinct mathematical objects, e.g. $\int f(x)\,dx$ for several anti-derivative operators (Riemann, Lebesgue,...).

[5] http://scholar.google.cz/scholar?q=Kovacik+Rakosnik

[6] http://search.mathweb.org/index.xhtml [7] http://searchpoint.ijs.si/

– Certain variations of notations are widely considered irrelevant, e.g. $\int f(x)\,dx$ and $\int f(y)\,dy$.

There are several math search systems and platforms available:

– *MathWebSearch*[8], by I. Şucan, M. Kohlhase (Bremen, GE);
– *MathDex*, by R. Miner et al. (Design Science, US) or *DLMF search*, A. Youssef (Washington, US);
– *EgoMath/Egothor*, J. Mišutka, L. Galamboš (Prague, CZ).

Other notable related work is:

– Mathematical formulae recognition from PDF, J. Baker, A. Sexton, V. Sorge, Birmingham, UK.
– Infty system, M. Suzuki, Kyushu, JP.
– ActiveMath web-based math-learning environment, P. Libbrecht, DKFI, Saarbrücken, GE.
– SWiM: A Semantic Wiki for Mathematical Knowledge Management, KWARC, Bremen, GE.

Math search system has to solve many technical aspects of search. In EgoMath system, these are e.g.

– normalization;
– linearization (search engine may work on strings/words);
– partial evaluation (e.g. distributivity);
– generalization (introduction of variables in the index) or
– ordering (for commutative operators).

Complexity of these issues are probably causing that there is not a widely used web search engine handling math yet.

> Automating the creation of useful digital libraries – that is, digital libraries affording searchable text and reusable output – is a complicated process, whether the original library is paper-based or already available in electronic form. – Simske and Lin [8]

4 Math Digital Libraries

There is the vision of the world-wide digital mathematics library [9].

We may classify levels of digital libraries of mathematics:

1.0 classical library + scanned bitmaps.
2.0 interconnected, crosslinked and validated repository of peer reviewed documents, possibly fully (not only metadata) indexed on the syntactic level.

[8] `http://www.mathweb.org/wiki/MathWebSearch`

3.0 dynamically personalized, formalized knowledge in rich semantic representation with logical inference and deduction.

Most DMLs today strive to attach rich metadata to the scanned page bitmaps (level one). The ideal 3.0 world remains as a vision for the next decades. There are attempts towards level 2.0 (DML-CZ, NUMDAM, Euclid[9]). Reference lists are considered as paper metadata and made available and linkable by current leading systems (as CrossRef[10]).

More and more applications can be build using the [richly tagged] paper full texts. One that is admired by users of digital library is application that provides links to similar papers ('see also' types of suggestions).

When the music changes, so does the dance. – African proverb

5 Math Paper Similarities

Showing similar papers functionality starts to be offered by several digital libraries and publisher. But how to find similar papers among other milions? How to evaluate the possible candidate lists? Which type of similarity is preferred?

We have tried to think about these kind of questions and did some experiments with the data of DML-CZ and NUMDAM. We have used bag of words vector models for paper representation, and computed similarities by three methods: *TFIDF* term weighting, *Latent Semantic Analysis* (LSI) and by *Random projections* [10]. They are available for author's evaluation on the DML-CZ web pages. We were stuck with evaluation, as almost no author was willing to go through computed lists of similar papers and to compare the results given by different methods. Top ordering comparisons done by experts was evaluated as too costly and unfeasible within the budget and time constraints. The only information available we can base the evaluation on are available metadata as MSC numbers, and article full texts. But another solution came to our mind: *random walking*.

Let us remind method of Larry Page to compute ranking of web pages [4]. Let $G = \langle N, L \rangle$ be a graph of interlinked documents and let $W_0[i,j] = 1$ iff there is link from node n_i to n_j. Let we define forward neighbours of a document as $F(i) = \{n_j | W_0[i,j] = 1\}$. Let we now row-normalize adjacency matrix of G: $W[i,j] = \frac{1}{|F(i)|}$ if $W_0[i,j] = 1$ and $W[i,j] = 0$ otherwise.

Page's algorithm takes row-normalized adjacency matrix $\boldsymbol{W}$ and vector $\boldsymbol{e}$ (internal source of score of n_i, constant across iterations) and iteratively computes

$$\boldsymbol{a}^{(k)} = \alpha \boldsymbol{a}^{(k-1)} \boldsymbol{W} + (1-\alpha)\boldsymbol{e}\,.$$

Resulting vector is $\boldsymbol{a} = \langle a_1, a_2, \ldots, a_{|N|} \rangle$, where a_i represents the 'score' (pagerank) of node n_i. For more information we refer to the original paper or to the recent application of it in the area of Natural Language Engineering [11].

[9] http://projecteuclid.org [10] http://crossref.org

Let now take one document of interest n_k, for which we want to compute the most similar ones. We think of forward neighbours set $F(i)$ as a *support of similarity* to the document n_k of interest, based on the 'local knowledge' of document n_i.

Vector $\boldsymbol{e}$ can be used for smoothing (all values set to $\frac{1}{|N|}$), or as a source of explicit knowledge. It may be plausible to set non zero values only to all documents sharing same Mathematical Subject Classification (MSC)[11] codes as the document of interest. After the (convergence) computation, vector $\boldsymbol{a}$ contains similarity-ranking of document of interest (and DL may expose links to the ten documents having highest similarity scores a_i).

This framework allows solution of different tasks: different F and $\boldsymbol{e}$ can be used to compute different kinds of similarity – *simtypes*. We think of

topicality: this simtype should find thematically closest papers. F may be based on some vector space document model (LSA), $\boldsymbol{e}$ may reflect common MSC.

plagiarity: F should be based on the number and length of common word or word synsets n-grams.

narrativity: narrative qualities are often neglected when computing document similarities. New ways of representing narrative qualities as Markov chain start to appear as in the recent paper by Hoencamp et al. [12]. F should be sent for documents with similar or same Markov chain.

or their weighted combinations.

Computation will be time-consuming though: convergence for every task (simtype) and *every document (node)* has to be computed. It is yet to be shown how it will work in practice and whether these 'vis maior' simtypes will be praised by [Eu]DML users.

6 Conclusion

In this paper, we have identified some specifics of mathematical documents and suggested solution to the similarities problem – how to find documents close to the given one using different definitions of similarity metric.

Acknowledgments. This work has been partially supported by the Academy of Sciences of Czech Republic under the project 1ET200190513 and by the Ministry of Education of CR within the Centre of basic research LC536 and National Research Programme 2C06009.

References

1. Sojka, P., ed.: Towards a Digital Mathematics Library. In Sojka, P., ed.: Proceedings of DML 2008, Birmingham, UK, Masaryk University (2008) `http://www.fi.muni.cz/~sojka/dml-2008-program.xhtml`.

[11] `http://www.ams.org/msc`

2. Sojka, P., ed.: Towards a Digital Mathematics Library. In Sojka, P., ed.: Proceedings of DML 2009, Grand Bend, Ontario, CA, Masaryk University (2009) http://www.fi.muni.cz/~sojka/dml-2009-program.html.
3. Sojka, P.: Digitization Workflow in the Czech Digital Mathematics Library. Math-for-Industry Lecture Note Series **22** (2009) 272–280.
4. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Seventh International World-Wide Web Conference (WWW 1998), Brisbane, Australia (1998).
5. Grigore, M., Wolska, M., Kohlhase, M.: Towards context-based disambiguation of mathematical expressions. Math-for-Industry Lecture Note Series **22** (2009) 262–271.
6. Kováčik, O., Rákosník, J.: On spaces $L_{p(x)}$ and $W_{k,p(x)}$. Czech Mathematical Journal **41** (1991) 592–618 http://dml.cz/handle/10338.dmlcz/102493.
7. Kohlhase, M., Sucan, I.: A Search Engine for Mathematical Formulae. In Calmet, J., Ida, T., Wang, D., eds.: AISC. Volume 4120 of Lecture Notes in Computer Science., Springer (2006) 241–253.
8. Simske, S.J., Lin, X.: Creating Digital Libraries: Content Generation and Re-Mastering. In: Proceedings of First International Workshop on Document Image Analysis for Libraries (DIAL 2004). (2004) pp. 33. http://doi.ieeecomputersociety.org/10.1109/DIAL.2004.1263235.
9. Jackson, A.: The Digital Mathematics Library. Notices Am. Math. Soc. **50**(4) (2003) 918–923.
10. Řehůřek, R., Sojka, P.: Automated Classification and Categorization of Mathematical Knowledge. In Autexier, S., Campbell, J., Rubio, J., Sorge, V., Suzuki, M., Wiedijk, F., eds.: Intelligent Computer Mathematics—Proceedings of 7th International Conference on Mathematical Knowledge Management MKM 2008. Volume 5144 of Lecture Notes in Computer Science LNCS/LNAI., Berlin, Heidelberg, Springer-Verlag (2008) 543–557.
11. Esuli, A., Sebastiani, F.: PageRanking WordNet Synsets: An Application to Opinion Mining. In: ACL, The Association for Computer Linguistics (2007) http://aclweb.org/anthology-new/P/P07/P07-1054.pdf.
12. Hoenkamp, E., Bruza, P., Song, D., Huang, Q.: An Effective Approach to Verbose Queries Using a Limited Dependencies Language Model. In Azzopardi, L., Kazai, G., Robertson, S.E., Rüger, S.M., Shokouhi, M., Song, D., Yilmaz, E., eds.: ICTIR. Volume 5766 of Lecture Notes in Computer Science., Springer (2009) 116–127 http://dx.doi.org/10.1007/978-3-642-04417-5_11.