December 19, 2018

Web search of mathematical expressions

Victor Cărbune

Thomas Deselaers

Web search of mathematical expressions

ABSTRACT

Internet search engines, which make information nearly universally accessible, still perform poorly at handling queries with mathematical content. The reasons for suboptimal handling of mathematical content by search engines include the following: the difficulty of entering mathematical expressions into search engines, the lack of standard encoding for mathematical content, the difficulty of determining if a given page or information source is relevant to a mathematical query, etc.

This disclosure describes techniques that address the aforementioned problems, e.g., by enabling easy entry of mathematical expressions into search engines, by creating a standard representation of mathematical content based on a machine-learning embedding of mathematical expressions, and by identifying pages relevant to a mathematical query using a machine-learned ranking system.

KEYWORDS

- Mathematical query
- Mathematical expression
- Search engine
- Machine learning
- Math search
- Equation search
- Embeddings

BACKGROUND

Internet search engines, which make information nearly universally accessible, still perform poorly at handling queries with mathematical content. The reasons for suboptimal handling of mathematical content by search engines include the following:

- difficulty of entering mathematical expressions into search engines;
- lack of standard encoding for mathematical content: although there are some ways to encode mathematical expressions, e.g., LaTeX, MathML, etc., there are no agreed-upon standards);
- difficulty of determining if a given page or information is relevant to a mathematical query; exact matching, e.g., by replacing variables or simplifying expressions, often make the matching problem harder; etc.

DESCRIPTION

This disclosure describes techniques that enable robust and user-friendly search of mathematical expressions. The search procedure is optimized towards mathematical expressions from end to end, e.g., from user interface to machine-learned page-ranking.

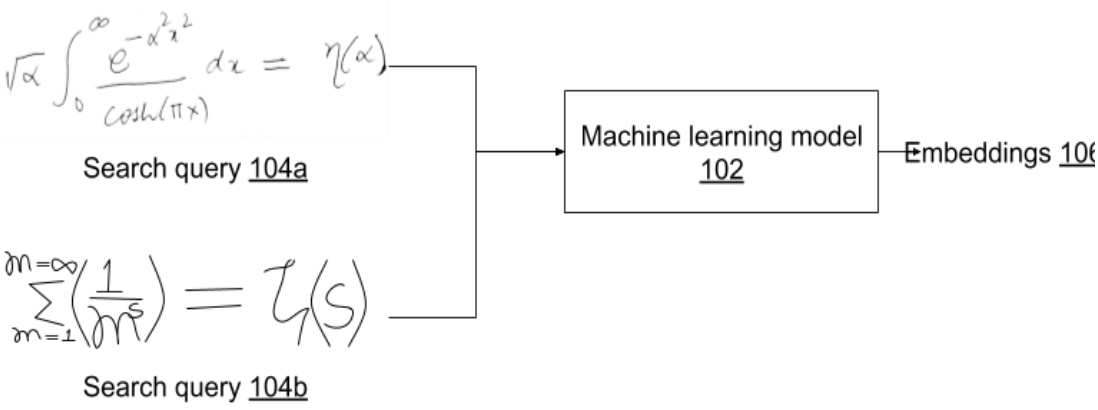

Fig. 1: Entering handwritten mathematical search queries

Entering mathematical search queries

Fig. 1 illustrates the entry of mathematical search queries, per techniques of this disclosure. A search engine accepts as input a picture (104a) of a mathematical expression or a handwritten form thereof (104b). A user can handwrite a mathematical expression, e.g., using their finger or stylus on a touch screen. The picture or handwritten mathematical expression is fed to a machine learning model (102), which produces embeddings (106) of the input mathematical expression. The embeddings serve as a representation of the input query for the purposes of searching. Additionally, the input expression, or its embedding, can be recognized and converted into a conventional machine-readable format, e.g., LaTeX or MathML. The embeddings representation serves as a natural standard format for mathematical expressions and is richer than conventional formats.

The machine learning model used to recognize the input mathematical search queries and develop embeddings thereof is implemented using, e.g., a neural network. Example types of neural networks that can be used include long short-term memory (LSTM) neural networks, recurrent neural networks, convolutional neural networks, etc. The machine learning model can also be, e.g., a generative machine learning model, a regression learning model, etc. Other machine learning models, e.g., support vector machines, random forests, boosted decision trees, etc., can also be used.

Standardizing formats for mathematical expressions

Indexed web pages that include mathematical expressions are standardized by using a machine learning model to generate embeddings of the web pages, e.g., the embeddings serve as a machine-readable standard representation of the rendered mathematical content of web

pages. Mathematical web pages are recognized and converted to the machine-readable format used by a search engine. This format is simply an embedding of one of the neural network layers that processes the image or handwritten equation.

Determining relevance of a web page to a given mathematical query

Once the mathematical content is in a standard format, the relevance of a web page to a given query is discovered by finding equations (or embeddings, e.g., standard representations) similar to the query within the page. This enables effective searching of a large fraction of mathematical queries, since many common equations have standard forms.

A web page is ranked for relevance to a given search query by using a machine-learning model to create and scored based on similarity between the embeddings of the given query and of mathematical equations on the web page. In this context, web pages are the sources of information that are to be searched. Such a technique to rank web-pages is termed as implicit ranking. Alternately, variable names in the search query and the web pages can be replaced with placeholders, and a scoring function can be defined that compares similarity between two equations. Such a technique for ranking web pages is termed as explicit ranking.

Explicit and implicit ranking can be combined to obtain more powerful matching signals. This enables fuzzy matching of equations, e.g., when searching for an equation using different characters than the ones in the web page, or when searching for a different form of an equation.

Training of machine learning models

The machine learning models described herein are trained by using similar equations written in different forms. The ranking system is trained to output a high-similarity score for

pairs of equations that are close together and a low score for unrelated ones. Synthetic data, e.g., as generated by equation editors or typesetters, can be used for such training.

The techniques for indexing of web content described herein apply to other content, e.g., videos, audios, etc., that include content having a mathematical or educational nature. Alternatively, or in addition to the techniques described above, heuristics relating to the metadata, e.g., annotations, around mathematical equations can be used as search signals. The techniques are also useful in other contexts, e.g., document storage and management software, special-purpose search engines such as those focused on scholarly content, etc.

In this manner, the techniques of this disclosure make mathematical content more accessible via web search and enable the testing for similarity of apparently unrelated equations.

CONCLUSION

This disclosure describes techniques that enable user-friendly and robust web search of mathematical expressions, e.g., by enabling easy entry of mathematical expressions into search engines, by creating a standard representation of mathematical content based on a machine-learning embedding of mathematical expressions, and by identifying pages relevant to a mathematical query using a machine-learned ranking system.

REFERENCES

[1] Zanibbi, Richard, Kenny Davila, Andrew Kane, and Frank Wm Tompa. "Multi-stage math formula search: Using appearance-based similarity metrics at scale." In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 145-154. ACM, 2016.