QuQn Map: *Qu*alitative-*Q*ua*n*titative Mapping of Scientific Papers

Xing Wang, Jason Lin, Ryan Vrecenar, Jyh-Charn Liu
Department of Computer Science and Engineering
Texas A&M University, College Station, TX 77843, USA
{senyalin, wxadb, jcliu}@tamu.edu

ABSTRACT

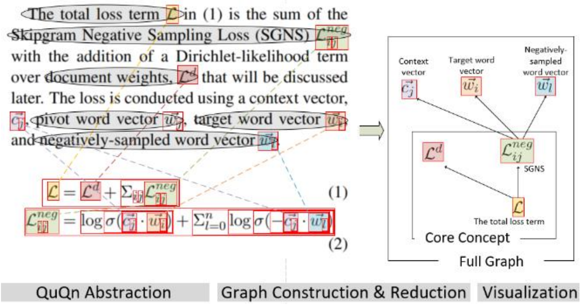
The Qualitative-Quantitative map (QuQn) is an abstraction of scientific papers based on automatically extracted mathematical expressions (ME) and their related words. MEs are widely used as the concise representation of physical concepts and the interactions among them. Mapping of dependency between MEs offers a succinct representation of the reasoning logic flow in a paper. By linking these MEs with adjacent plaintexts gives insights on their qualitative descriptions. By analyzing the dependency relationship among MEs, one can prune QuQn nodes by different filters, such as duplication removal, labeling of source / intermediate / concluding nodes, and node connectivity. The end goal is to assist the reader to more efficiently capture the essence of technical contents from the extracted structural relationship. A visualization tool prototype is developed to support interactive browsing of the technical contents at different granularities of detail.

KEYWORDS

QuQn map, mathematical expression, visualization,

1 INTRODUCTION

Technical writing compiles complex mathematical system into a linear sequence. To digest the original idea, one must walk through forward and backward to understand the complex relations, as well as look up the external materials. It's a challenging task for normal readers to track the notations and different thread of discussion. But to design automated algorithms to build the logic flow, there is another gap between the semantics and the layout placement of the characters for plaintext words and mathematical expressions (ME). To overcome the information overload and semantical gap challenge, we propose the concept of Qualitative-Quantitative (QuQn) map abstraction to help readers quickly grasp the main idea and dive into detail when necessary. The qualitative aspect links ME to their physical/abstraction concept and quantitative show the interconnection between MEs based on dependency analysis after converting all MEs into the same semantical taxonomy. Since the human visual system is the most powerful perception for information understanding, we visualize the QuQn map with customized spatial layout and color style to highlight the dependency relationship. The QuQn map is progressively pruned varying from the full Graph to the core concept.



The total loss term $\mathcal{L}$ in (1) is the sum of the Skipgram Negative Sampling Loss (SGNS) $\mathcal{L}^{neg}_{ij}$ with the addition of a Dirichlet-likelihood term over document weights, $\mathcal{L}^{d}$ that will be discussed later. The loss is conducted using a context vector, $\vec{c_j}$, pivot word vector $\vec{w_j}$, target word vector $\vec{w_i}$, and negatively-sampled word vector $\vec{w_l}$.

$$\mathcal{L} = \mathcal{L}^{d} + \Sigma_{ij}\mathcal{L}^{neg}_{ij} \quad (1)$$

$$\mathcal{L}^{neg}_{ij} = \log\sigma(\vec{c_j}\cdot\vec{w_i}) + \Sigma^{n}_{l=0}\log\sigma(-\vec{c_j}\cdot\vec{w_l}) \quad (2)$$

Figure. 1. Tri-layer QuQn map architecture

The QuQn system overcomes the ad-hoc normalization of ME and linking by introducing three layers of abstraction shown in Fig. 1: QuQn abstraction, progressive graph construction, and visualization. The QuQn abstraction is designed to contain the essence of a paper from the qualitative and quantitative aspects. First, for all MEs in a paper, we extract, parse and convert all MEs into the same mathematical semantical taxonomy, no matter the ME's original form as Latex, representational MathML, content MathML or even ME analyzed from PDF files. MEs is also decomposed into sub-expression when possible shown as the nested red rectangle box. Second, the denotation of ME as other MEs or plaintext phrase are extraction to build the dependency relationship among MEs as well as link the ME with qualitative plaintext marked in the grey ellipse. Then we progressively prune the dependency relationships based on different requirements. At last, we design a visual program [1] to present different layers of information in a limited 2D space using various visual factors: vertical position to differentiate hierarchical layers and the colors to differentiate the components. In this way, readers could quickly grasp the main skeleton of an idea and follow the inference process.

The rest of the paper is organized as follows: In section 2, we present the QuQn abstraction with an emphasis on the introduction of the semantical taxonomy of ME and ME denotation. In section 3, we offer a case study to show the construction of the skeleton graph based on the QuQn abstraction; In section 4, the visualization design is introduced; At last, we present related work and conclude our work.

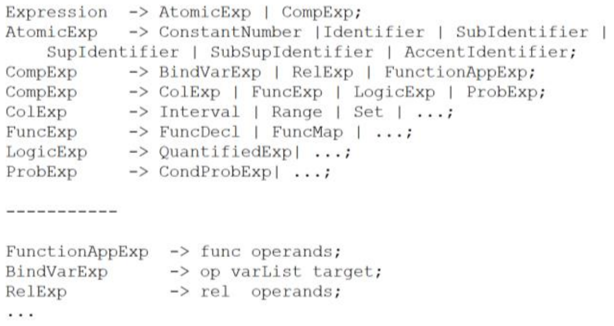
2 QuQn Abstraction

The QuQn map is represented by a pair $\langle\mathcal{M},\mathcal{D}\rangle$, where $\mathcal{M} = \{m_i\}$ as all the ME and sub expression in a document and $\mathcal{D}$ are set of ME denotation either as a plaintext description or other equivalent ME. Since the extraction of equivalent ME or related ME requires the semantical level understanding of ME, we first introduced our semantical taxonomy of ME and define useful concepts such as “equal”, “sub-component”, and “left hand side”. Then we present the formulation of ME denotation, which play important role in linking MEs and associate ME with plaintext.

2.1 ME Processing

We will first brief the semantical taxonomy of ME type and the composition of each type of ME. Then the concept of equal, subcomponent and left/right-hand side are introduced for the discussion for relation extraction between MEs.

2.1.1 Semantical taxonomy of ME



```
Expression   -> AtomicExp | CompExp;
AtomicExp    -> ConstantNumber |Identifier | SubIdentifier |
    SupIdentifier | SubSupIdentifier | AccentIdentifier;
CompExp      -> BindVarExp | RelExp | FunctionAppExp;
CompExp      -> ColExp | FuncExp | LogicExp | ProbExp;
ColExp       -> Interval | Range | Set | ...;
FuncExp      -> FuncDecl | FuncMap | ...;
LogicExp     -> QuantifiedExp| ...;
ProbExp      -> CondProbExp| ...;

-----------

FunctionAppExp  -> func operands;
BindVarExp      -> op varList target;
RelExp          -> rel  operands;
...
```

Figure 2: Semantical Taxonomy of ME & ME composition

Due to the flexibility of expressing the same semantical concept in Latex/MathML, tedious *ad-hoc* normalization procedures are required to eliminate the representation difference. In this work, we proposed to convert all representation-level ME into the semantical level by introducing the semantical taxonomy of ME and ME composition. Our semantical ME taxonomy is mostly built on top of the Content MathML standard [13], with extension to meet the need for special fields. The hierarchy and composition relationships among different ME types are is shown as grammar in Fig.2.

MEs are organized into atomic expressions and compounded expressions. The atomic expressions include constant, and identifier with optional subscript/superscript/accent. The compounded expressions include the relation expression, the function application expression, and the binding variable expression. The function concept is a generalized concept that also include the normal operation such as plus and multiply. Besides, the compounded expressions also include domain specific expression, such as the ColExp (set theory), FuncExp (functional analysis), LogicExp (logic) and ProbExp (probability). The grammar could be expanded and loaded dynamically as needed.

Besides the hierarchy of taxonomy, each semantical ME class also has its own subcomponents. For example, a function application expression consists of the function, and the arguments to be applied on. A complete grammar could be found in [2].

2.1.2 Operation/Relation of MEs. We will explain a few important concepts defined for MEs based on the example in Fig. 3.

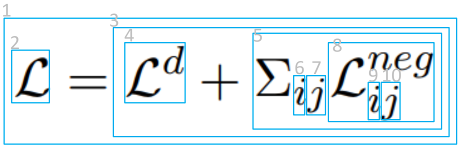

1
2
3
4
5
6
7
8
9
10
$\mathcal{L} = \mathcal{L}^d + \Sigma_{ij} \mathcal{L}_{ij}^{neg}$

Figure 3: Illustration of an ME example

In Fig. 3, we have an ME as $\mathcal{L} = \mathcal{L}^d + \sum_{ij} \mathcal{L}_{ij}^{neg}$, which consist of a total of 10 (sub)expression. It’s composed of direct subexpression: $\mathcal{L}$, $\mathcal{L}^d$, and $\sum_{ij} \mathcal{L}_{ij}^{neg}$, where $\sum_{ij} \mathcal{L}_{ij}^{neg}$ is further composed of i, j, and $\mathcal{L}_{ij}^{neg}$. For an ME m_i, we denote $\Phi(m_i)$ as its direct subcomponents, and $\Psi(m_i)$ as all its subcomponents. Note that the same expression might occur multiple times within a document, e.g., m_6 and m_9 are all identifier “i”.

We have the following three concepts for ME denotation and ME relation extraction:

- $m_i = m_j$ if the two expressions are the same, such as the case of m_6 and m_9.
- The ME m_i is a subcomponent of m_j, denoted as $m_i \in m_j$, iff $\exists m_{i'} \in \Psi(m_j), m_i = m_{i'}$.
- For some ME types such as relation expression, and function declaration expression, we could define the left(right)-hand side function $L(R)HS$. For the example above, $LHS(m_1) = m_2$. If there is no valid LHS, $LHS(m) = null$.

2.1.3 ME Extraction and Parsing. Based on the semantical taxonomy defined in 2.2.1, we build an in-house system that could parse either LateXML [3] or the directly extract and parse MEs from the PDF files [2][4] into the semantical taxonomy in Fig. 2.

2.2 Denotation Extraction

Denotation refers to the semantical equivalent information for an ME, which is critical to build relations between MEs and link an ME to the plaintext concepts. A denotation could be plaintext definition from qualitative aspect; it could also be another ME equals the ME under concern from the quantitative aspect.

We will first discuss the denotation defined by equivalent ME (*ME denotation*), given an ME m with LHS and RHS with the relation “=” or assignment, we could build a denotation $\langle LHS(m), \mathrm{RHS}(m) \rangle$. We denote the set of ME denotation as $\mathcal{D}_M$. Besides the relation between different mathematical expressions, each ME m_i is optionally associated with the plaintext definition in the neighbor context, which composes of a sequence of words $W_i = \{w_i^j\}$. For ME with definition, we build *plaintext denotation* as $\langle m, W \rangle$. The full set of plaintext denotation is denoted as $\mathcal{D}_P$. We implemented a rule-based ME definition extraction system using the rules defined in [5], which could achieve the state of the art performance as machine learning based methods [5][6] based on the evaluation dataset for the math understanding task in NTCIR [7].

3 Skeleton Graph Construction and Pruning

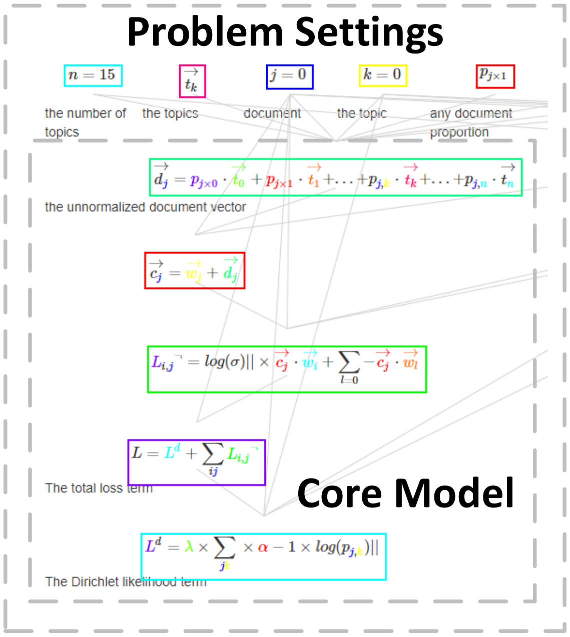

Problem Settings
$n = 15$
$\vec{t_k}$
$j = 0$
$k = 0$
$p_{j\times 1}$
the number of topics
the topics
document
the topic
any document proportion
$\vec{d_j} = p_{j\times 0} \cdot \vec{t_0} + p_{j\times 1} \cdot \vec{t_1} + \ldots + p_{j,k} \cdot \vec{t_k} + \ldots + p_{j,n} \cdot \vec{t_n}$
the unnormalized document vector
$\vec{c_j} = \vec{w_j} + \vec{d_j}$
$L_{i,j}{}^{\neg} = log(\sigma)|| \times \vec{c_j} \cdot \vec{w_i} + \sum_{l=0} - \vec{c_j} \cdot \vec{w_l}$
$L = L^d + \sum_{ij} L_{i,j}{}^{\neg}$
The total loss term
Core Model
$L^d = \lambda \times \sum_{jk} \times \alpha - 1 \times log(p_{j,k})||$
The Dirichlet likelihood term

Figure 4: Colored visualization of the skeleton graph pruned from QuQn map

The QuQn map covers the essence of a publication, but too much information such as repetitive mentioning and low-level details. Researchers read abstraction and introduction as a pre-filter for relevant papers. Similarly, we propose the skeleton graph, a progressive visualization of publications. An example of the skeleton graph is show in Fig. 4 for the paper *lda2vec* [8]. Different level of pruning could be applied. A basic duplication removal pruning results in a succinct representation including the problem settings and the derivation for high-level concept. A further pruning of the primary problem settings leads to the core discussion, which is usually the main contribution. In the skeleton graph, each ME is surrounded by a bounding box with the assigned color, and optional plaintext definition are placed under the box. The dependency relationship along MEs are shown as grey edges, and each sub-expression is colored accordingly. The visualization design will be discussed in section 4.

The skeleton graph is pruned from the QuQn map based on two criteria. The first criterion is to keep the MEs with denotation only for users to understand the semantics of every ME node. The second criterion is to remove duplicate occurrence of the same ME to clear redundancy. To meet the criterion 1, we only keep the ME with denotations, $\mathcal{M}_d = \{m: \langle m, * \rangle \in \mathcal{D}_M \cup \mathcal{D}_P\}$. Secondly, we only keep the first occurrence of the ME to remove duplication. Formally, among the MEs with denotation $\mathcal{M}_d = \{m_{d_i}\}$, we remove the ME with multiple denotation and get $\mathcal{M}_d^r$. This is equivalent to say if $m_{d_i}, m_{d_j} \in \mathcal{M}_d$, and $m_{d_i} = m_{d_j}$, then we only keep the m_{d_i} in the ME with denotation. After this process, we get the dependency graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{E} \rangle$ with the complete dependency information, where $\mathcal{N} = \mathcal{M}_d^r$, $\mathcal{E} = \{\langle m_i, m_j \rangle : \langle m_i, m' \rangle \in \mathcal{D}_M, m_j \in \Psi(m'), m_i, m_j \in \mathcal{M}_d^r\}$. The condition $m_i, m_j \in \mathcal{M}_d^r$ requires that the nodes are in the reduced set. The condition $\langle m_i, m' \rangle \in \mathcal{D}_M, m_j \in \Psi(m')$ states that m_j is the subexpression of m', which is equivalent to m_i.

But the dependency graph above is still too crowded for visualization. A series of post-processing is conducted to make the graph less overloaded.

- Remove indirect edges. If there exist edge $\langle m_i, m_j \rangle$, $\langle m_j, m_k \rangle$, we will also have edge $\langle m_i, m_k \rangle$, which is indirect because of the intermediate node m_j. They will not add new information and make the graph overloaded.
- Choose the largest connected components. Some local discussions not connected to the main logic flow will be removed.
- Remove problem settings. The graph might still be quite crowed with lots of MEs for the problem settings as shown in Fig. 4. We detected these primary MEs for problem settings based on the following rules. The primary ME includes: 1) identifiers (with optional sub/superscripts, accent) which do not interact with each other and 2) relation expression with a constant number on the right-hand side, which are usually the detail of the implementation. When the primary concepts are removed, only the core contribution of this paper will be shown.

4 Graph Visualization

Given the graph constructed and reduced, the next task is to visualize the dependency graph so that users could quickly identify the essential elements and understand the detail. Following the visual program concept proposed by Tufte [1], we will use the spatial and color within a 2D space to make readers quickly identify the dependency relationships among MEs.

Spatially, we group the MEs into layers based on the depth of the ME in the dependency tree. Since a recursive definition is rarely allowed in scientific elaboration, we remove some edges to make the graph acyclic as a tree, in which the large ME composing of the smaller MEs. The depth of each tree node is a good indicator how coarse vs. detail the ME is. The larger the depth is, the coarser the concept is. Further, grouping the MEs into layers could reduce the possibility of crossing edges so that the graph is easier to follow. But, there might be too many MEs in one layer that could not be fit into one line in the limited 2D space, we break such layer into multiple lines.

From the aspect of color, we use the same color c_i for the same ME m_i across the whole filtered graph $\mathcal{G}$ to help people quickly identify the same ME in different places. The color c_i will also be used as the color of bounding box for m_i to indicate the first time the ME m_i is presented. Further, the hue contrast of neighbor MEs is maximized to enhance the differentiability for ease of identification. The key challenge is that we only have limited color. We use the color wheel concept and choose the color that is distant from existing color with smallest standard deviation to enhance the contrast in a limited number of color choice. For example, if an ME only have one neighbor of color blue-violet, the ME will be assigned color yellow-orange to maximize the contrast. If an ME is with two neighbor of color red-orange and blue-violet, the ME will be assigned color yellow-green to maximize the total contrast as well as minimize the standard deviation of contrast.

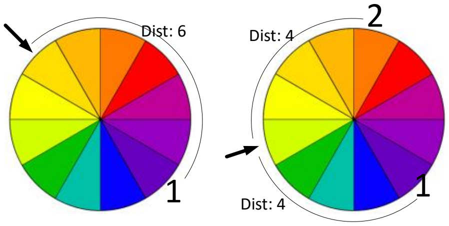


Figure. 5. the color wheel and color choosing algorithm

We draw the skeleton graph on the canvas of HTML5. The MathJax javascript library renders the math in Latex representation, where the color of sub ME is encoded using the \color command in Latex.

5 RELATED WORK

Recovering the mathematical logic of publication has been studied at coarse and detail level. At the coarse level, math centered publication analysis tried to extract the math component such as definition, theorem, lemma, etc. and extract the dependency among the blocks [9][10]. They are mostly based on string matching and rule-based relation classification. But the coarse-level mathematical structures failed to capture many important elements such as the MEs and interaction between ME and ME/plaintext.

At the detail level, a few studies [11][12] used the dependency relationship among MEs to transfer the semantics from subexpression for improving the searching performance of mathematical information retrieval. There are three aspects of limitation of existing work: limited dependency, normalization, no information reduction for visualization. For the last issue, non-polished dependency graph will contain too much information for end-users to consume, and we propose a pruning procedure to reduce the graph from detail into coarse progressively. The first two issues are both limited by the representation markup language. Though there is content MathML [13] standard proposed, Latex and representation MathML is still the most common way people input math, which only contains the representation layout information. The challenge from the Latex and representation MathML is that the same ME could be written in different ways. Normalization of different mathematical expression representation is one crucial step for better information retrieval [14]. Heuristics are proposed on the MathML level [11], removal of structures (mrow, parentheses, attachment, right-hand side ME), and case normalization. After normalization, the ME dependency graph is mostly constructed over the representational MathML based on string matching or subexpression matching [11]. These approaches highly depend on the quality of MathML and will miss dependency at the semantical level. In this paper, we try to alleviate the first two issues by converting all ME into the same semantical taxonomy. Since PDF is the de facto standard for publishing, our system also has a pipeline to extract and parse ME from the PDF files.

Our skeleton graph is a type of analytical product to assist reading. Some related project includes Utopia [14], Math-vis [15], and Arixv-Vanity[16]. Utopia to enhance the reading experience of the medical domain by matching external resources such as terminology dictionary and reference. As for math center publication visualization, the dependency graph in [11] are not designed for human to read. Math-vis is limited to visualize a single MathML and differential analysis of a pair of MathML. Vanity [16] is an excellent project that is built on top of LateXML to show Arxiv papers on HTML page without content analysis for the MathML.

6 CONCLUSION

This paper presents a tri-layer Qualitative-Quantitative map abstraction to describe the technical essence of scientific papers. The QuQn map is novel in that all ME are encoded based on a semantical taxonomy to avoid the error-prone and ad-hoc normalization. On top of the QuQn abstraction, we have modularized design so that it could easily be expanded to progressively show different levels of information from the coarsest skeleton graph to the most detail overlays on the original documents. The case study shows that the QuQn map could capture the global idea of paper and the detail elaboration.

REFERENCES

[1] Tufte, Edward R. "The visual display of quantitative information." Journal for Healthcare Quality 7.3 (1985): 15.

[2] Wang, Xing, and Jyh-Charn Liu. "A content-constrained spatial (CCS) model for layout analysis of mathematical expressions." ICDIM, 2017.

[3] Miller, Bruce. "LaTeXML: ALatex to xml converter." Web Manual at http://dlmf. nist. gov/LaTeXML/, seen September2007 (2010).

[4] Xing Wang, Jyh-Charn Liu, "On a Font Setting based Bayesian Model to extract mathematical expression in PDF files", ICDAR 2017

[5] Kristianto, Giovanni Yoko, et al. "Extracting definitions of mathematical expressions in scientific papers." Proc. of the 26th JSAI. 2012.

[6] Pagael, Robert, and Moritz Schubotz. "Mathematical language processing project." arXiv preprint arXiv:1407.0167 (2014).

[7] Aizawa, Akiko, Michael Kohlhase, and Iadh Ounis. "NTCIR-10 Math Pilot Task Overview." In NTCIR. 2013.

[8] Moody, Christopher E. "Mixing dirichlet topic models and word embeddings to make lda2vec." arXiv preprint arXiv:1605.02019 (2016).

[9] Nakagawa, Koji, Akihiro Nomura, and Masakazu Suzuki. "Extraction of logical structure from articles in mathematics." International Conference on Mathematical Knowledge Management. Springer, Berlin, Heidelberg, 2004.

[10] Solovyev, Valery, and Nikita Zhiltsov. "Logical structure analysis of scientific publications in mathematics." WIMS. ACM, 2011.

[11] Kristianto, Giovanni Yoko, Goran Topić, and Akiko Aizawa. "Utilizing dependency relationships between math expressions in math IR." Information Retrieval Journal 20.2 (2017): 132-167.

[12] Krenn, Mario, et al. "Automated search for new quantum experiments." Physical review letters 116.9 (2016): 090405.

[13] Ausbrooks, Ron, et al. "Mathematical markup language (MathML) version 2.0 . W3C Recommendation." World Wide Web Consortium 2003 (2003).

[13] Ružicka, Michal, Petr Sojka, and Martin Líška. "Math indexer and searcher under the hood: History and development of a winning strategy." Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. 2014.

[14] Attwood, Terri K., et al. "Utopia documents: linking scholarly literature with research data." Bioinformatics 26.18 (2010): i568-i574.

[15] Schubotz, Moritz, Norman Meuschke, Thomas Hepp, Howard S. Cohl, and Bela Gipp. "VMEXT: A Visualization Tool for Mathematical Expression Trees." In International Conference on Intelligent Computer Mathematics, pp. 340-355. Springer, Cham, 2017.

[16] https://www.arxiv-vanity.com/