Logical Structure Analysis of Scientific Publications in Mathematics

Valery Solovyev, Nikita Zhiltsov

Kazan (Volga Region) Federal University, Russia

WIMS'11
Overview

- LOD Cloud has been growing at 200-300% per year since 2007*
- Prevalent domains: government (43%), geographic (22%) and life sciences (9%)
- However, it lacks data sets related to academic mathematics

*C.Bizer et al. State of the Web of Data. LDOW WWW’11
Background

Proposed Semantic Model

Analysis Methods

Experiments and Evaluation

Prototype
Mathematical Scholarly Papers

Essential features

- Well-structured documents
- The presence of mathematical formulae
- Peculiar vocabulary ("mathematical vernacular")
Research Objectives

Current study
- Specification of the document logical structure
- Methods for extracting structural elements

Long-term goals
- A large corpus of semantically annotated papers
- Semantic search of mathematical papers
Modelling the Structure of Scientific Publications

ABCDE format

- LaTeX-based format to represent the narrative structure of proceedings and workshop contributions
- Sections:
  - Annotations (Dublin Core metadata)
  - Background (e.g. description of research positioning)
  - Contribution (description of the presented work)
  - Discussion (e.g. comparison with other work)
  - Entities (citations)
Modelling the Structure of Scientific Publications

SALT

- LaTeX-based authoring tool for generating semantically annotated PDF documents
- Three ontologies:
  - SALT Document Ontology
  - SALT Annotation Ontology
  - SALT Rhetorical Ontology
SALT Layers
Mathematical Knowledge Representation

- Languages for formalized mathematics
  - Mizar
  - Coq
  - Isabelle

- Semiformal math languages
  - HELM ontology
  - MathLang
  - OMDoc format (+ OMDoc ontology, sTeX)

- Presentation/authoring formats
  - PDF
  - \LaTeX
Mathematical Knowledge Representation

trade-off?
Trade-off Candidates

- arXMLiv format
  - XHTML+MathML
  - Marked up theorem-like elements, sections, equations
  - Automatic conversion for LaTeX documents with styles of available bindings (LaTeXML)
  - 60% of arXiv.org were converted into the format

- Present work
  - Follow the slides ⇒
1 Background

2 Proposed Semantic Model

3 Analysis Methods

4 Experiments and Evaluation

5 Prototype
Mathematical Knowledge Representation
Proposed Semantic Model

- It is an ontology that captures the structural layout of mathematical scholarly papers (as in the LaTeX markup)
- The segment represents the finest level of granularity and has the properties:
  - starting and ending positions
  - the text or math contents
  - functional role
- Select most frequent segments from sample collections of genuine papers
- Consider synonyms as one concept (e.g. conjecture and hypothesis)
Proposed Semantic Model (cont.)

- Select **basic semantic relations** between segments from the prior-art models
- Integration with SALT Document Ontology classes:
  - Publication
  - Section
  - Figure
  - Table
Ontology Elements
http://cll.niimm.ksu.ru/ontologies/mocassin#
1 Background

2 Proposed Semantic Model

3 Analysis Methods

4 Experiments and Evaluation

5 Prototype
Logical Structure Analysis

- The ontology specifies a controlled vocabulary to semantic analysis
- Two analysis tasks:
  - recognizing the types of document segments
  - recognizing the semantic relations between them
Example (cont.)
Example (cont.)
Recognizing the Types of Document Segments

We exploit the LaTeX markup extensively

1. Elicit a LaTeX environment
2. Associate it with a string that may be
   either the environment name
   or the environment title (if available)
3. Filter out standard formatting environments (e.g. center, align, itemize)
4. Compute string similarity between a string and canonical names of ontology concepts
5. Check if the found most similar concept is appropriate using a predefined threshold
Recognizing Navigational Relations

The *dependsOn* and *refersTo* relations are navigational.

**Assumption**

Navigational relations are induced by *referential* sentences.

**Examples**

- “By applying Lemma 1, we obtain …” (*dependsOn*)
- “Theorem 2 provides an explicit algorithm …” (*refersTo*)
Recognizing Navigational Relations

Supervised method

1. Given a segment $S$; split its text into sentences, tokenize and do POS tagging

2. **Referential sentences** are ones that contain the `\ref` command entries

3. For each sentence:
   - find mentioned segments; each of them makes a pair with $S$ (**type feature**)
   - for each pair, compute relative positions of segments normalized by the document size (**distance feature**)
   - build a boolean vector for its verbs (**verb feature**)
Recognizing Navigational Relations (cont.)

Supervised method

Example training instance

<table>
<thead>
<tr>
<th>t1</th>
<th>t2</th>
<th>d1</th>
<th>d2</th>
<th>add</th>
<th>...</th>
<th>apply</th>
<th>...</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>proof</td>
<td>lemma</td>
<td>0.09</td>
<td>0.27</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>dependsOn</td>
</tr>
</tbody>
</table>

- Train a learning model using these features and a labeled example set
- Apply the model to classify new induced relations
Recognizing Restricted Relations

The `hasConsequence`, exemplifies and proves relations are restricted

Assumption

Restricted relations occur between consecutive segments
According to the ontology, restricted relations involve instances of three types, separately: *Corollary*, *Example* and *Proof*

1. Seek a segment of one of these types
2. Find its segments-predecessors
3. Filter out segments of inappropriate types
4. Return the closest predecessor
1 Background

2 Proposed Semantic Model

3 Analysis Methods

4 Experiments and Evaluation

5 Prototype
Experimental Setup

Collections

- 1355 papers of the “Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika” journal
- A sample of 1031 papers from arXiv.org

Implementation

An open source Java library built upon:

- LaTeX-to-XML converters
- GATE framework
- Weka
- Jena

See http://code.google.com/p/mocassin
Segment Recognition Evaluation

- Evaluation on the arXiv sample only
- Q-gram string matching algorithm was used
- The threshold value was optimized w.r.t. $F_1$-score

<table>
<thead>
<tr>
<th>Type</th>
<th># of true instances</th>
<th>$F_1$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axiom</td>
<td>5</td>
<td>1.000</td>
</tr>
<tr>
<td>Claim</td>
<td>114</td>
<td>0.987</td>
</tr>
<tr>
<td>Conjecture</td>
<td>152</td>
<td>0.987</td>
</tr>
<tr>
<td>Corollary</td>
<td>1715</td>
<td>0.995</td>
</tr>
<tr>
<td>Definition</td>
<td>1838</td>
<td>1.000</td>
</tr>
<tr>
<td>Example</td>
<td>771</td>
<td>0.999</td>
</tr>
<tr>
<td>Lemma</td>
<td>4061</td>
<td>0.998</td>
</tr>
<tr>
<td>Proof</td>
<td>4943</td>
<td>0.997</td>
</tr>
<tr>
<td>Proposition</td>
<td>3052</td>
<td>0.999</td>
</tr>
<tr>
<td>Remark</td>
<td>2114</td>
<td>1.000</td>
</tr>
<tr>
<td>Theorem</td>
<td>4670</td>
<td>0.991</td>
</tr>
<tr>
<td>other</td>
<td>671</td>
<td>0.892</td>
</tr>
</tbody>
</table>
Ontology Coverage Evaluation

- Evaluation on the both entire collections (“Izvestiya” and arXiv)
- **Equations** are most ubiquitous segments (52% and 69%, respectively)
- The ontology covers types of 91.9% and 91.6% of segments (with SALT Section class – 99.5% and 99.6%)
Distribution of Segment Types

The graph shows the percentage of segment occurrences for different types of segments in the documents Izvestiya and arXiv. The x-axis represents the different types of segments (Theorem, Proof, Lemma, Remark, Corollary, Definition, Proposition, Example, others, Claim, Conjecture), and the y-axis represents the percentage of segment occurrences.

- **Theorem**: Approximately 30% in Izvestiya, about 15% in arXiv.
- **Proof**: Approximately 15% in Izvestiya, about 10% in arXiv.
- **Lemma**: Approximately 20% in Izvestiya, about 10% in arXiv.
- **Remark**: Approximately 15% in Izvestiya, about 5% in arXiv.
- **Corollary**: Approximately 10% in Izvestiya, about 5% in arXiv.
- **Definition**: Approximately 10% in Izvestiya, about 5% in arXiv.
- **Proposition**: Approximately 5% in Izvestiya, about 5% in arXiv.
- **Example**: Approximately 5% in Izvestiya, about 2.5% in arXiv.
- **Others**: Approximately 5% in Izvestiya, about 2.5% in arXiv.
- **Claim**: Approximately 2.5% in Izvestiya, about 1% in arXiv.
- **Conjecture**: Approximately 2.5% in Izvestiya, about 1% in arXiv.

The graph indicates that Theorem and Proof segments are the most frequent in Izvestiya, whereas in arXiv, Lemma and Remark segments are more common.
Evaluation of Navigational Relation Recognition

- A paper contains 51.4 (Izvestiya) and 53.9 (arXiv) referential sentences on the average
- 243 referential sentences were randomly selected and manually annotated
- 95% were true navigational relations
- A decision tree learner (C4.5) was trained
- The results were from 10-fold cross validation

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>$F_1$-score refersTo</th>
<th>$F_1$-score dependsOn</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td>0.663</td>
<td>0.566</td>
<td>0.752</td>
</tr>
<tr>
<td>type+distance</td>
<td>0.658</td>
<td>0.663</td>
<td>0.704</td>
</tr>
<tr>
<td>type+verb</td>
<td>0.704</td>
<td>0.653</td>
<td>0.770</td>
</tr>
<tr>
<td>type + distance + verb</td>
<td>0.741</td>
<td>0.744</td>
<td>0.772</td>
</tr>
</tbody>
</table>
A Cloud of Frequent Verbs

use, estim, prove, obtai, satisfi, defini, appli, support, appli, write, hold, let, establish, give, complet, show, discuss, exist, allow, split, mean, integr, note, point, claim, restrict, photo, make, increas, estab, given, provid, follow, have, thank, discuss, allow, given, show, discuss, exist, allow, split, mean, integr, note, point, claim, restrict, photo, make, increas, estab
Evaluation of Restricted Relation Recognition

- Evaluation on the arXiv sample only
- 10% of the documents which contain certain segments were randomly selected
- For each such a segment, corresponding relations were annotated manually
- Known issues: imported corollaries and examples for arbitrary text fragments

<table>
<thead>
<tr>
<th>Relation</th>
<th># of instances</th>
<th>$F_1$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>hasConsequence</td>
<td>178</td>
<td>0.687</td>
</tr>
<tr>
<td>exemplifies</td>
<td>62</td>
<td>0.613</td>
</tr>
<tr>
<td>proves</td>
<td>216</td>
<td>0.954</td>
</tr>
</tbody>
</table>
Conclusion on Evaluation

- The ontology covers the largest part of the logical structure and appears to be feasible for automatic extraction methods.
- The task of segment type recognition has been accomplished.
- The method for recognizing navigational relations establishes ground truth, however, a large-scale evaluation and learning model selection are required.
- The baseline method for recognizing restricted relations must be improved by leveraging additional information (discussed in the paper!)
1 Background

2 Proposed Semantic Model

3 Analysis Methods

4 Experiments and Evaluation

5 Prototype
Prototype

A prototype:

- demonstrates our ongoing research on semantic search of mathematical papers
- incorporates the logical structure analysis methods
- is integrated with arXiv API
- enables enhanced search for arXiv papers and visualization of their logical structure
- publishes the semantic index as Linked Data via SPARQL endpoint
Formulating a Query

http://cll.niimm.ksu.ru/mocassin
A strengthening of the Nyman-Beurling criterion for the Riemann hypothesis. 2 (Proof)
Luiz Duarte

\[ c = \lim_{n \to \infty} \left( 1 + \frac{1}{n} \right) \]
ON NYMAN-BEURLING CRITERION

Proof. We shall only sketch the proof of this proposition. Applying the Fourier-Mellin map (2.9) to \( f_{\rho} + \chi \) we have from Plancherel's theorem that

\[
2\pi \| f_{\rho} + \chi \|_2^2 = \int_{\mathbb{R}(c) = 1/2} \left| \zeta(z) \sum_{a=1}^{n} \frac{\mu(a)}{a^{s+c}} - 1 \right|^2 \left\| \frac{dz}{|z|^2} \right\|^2
\]

\leq 2 \int_{\mathbb{R}(c) = 1/2} \left| \zeta(z) \left( \sum_{a=1}^{n} \frac{\mu(a)}{a^{s+c}} - \frac{1}{\zeta(z+c)} \right) \right|^2 \left\| \frac{dz}{|z|^2} \right\|^2

+ 2 \int_{\mathbb{R}(c) = 1/2} \left| \frac{\zeta(c)}{\zeta(z+c)} - 1 \right|^2 \left\| \frac{dz}{|z|^2} \right\|^2.

The second integral on the right-hand side above is estimated to be \( \ll c^{3/4} \) as follows. If the distance between \( z = \frac{1}{2} + it \) and the nearest zero of \( \zeta \) is larger than \( \delta \), say, the upper bound

\[
\left| \frac{\zeta(z)}{\zeta(z+c)} - 1 \right| \ll \delta^{-1} (|t| + 1)^{1/4}
\]

follows from the classical estimate

\[
\frac{\zeta(s)}{\zeta(s)} = \sum_{|\tau| = 1} \frac{1}{s - \rho} + O(\log(2 + |s|)),
\]

where \( s = \sigma + it \), \( 1/2 \leq \sigma \leq 3/4 \), \( \tau \in \mathbb{R} \) and \( \rho = \beta + i\gamma \) denotes a generic zero of the \( \zeta \) function. By integration and exponentiation, provided \( \tau/\delta \) is small enough. In the other case, one uses an estimate of Burnol stating that under the Riemann hypothesis...
Summary

- The proposed approach aims to analyze the structure of mathematical scholarly papers in an automatic way
- Our ontology provides a controlled vocabulary for analysis
- The methods elicit document segments in terms of the ontology
- The extracted semantic graph can be used for:
  - discovering important document parts
  - semantic search of theoretical results
Thanks for your attention!
Questions?