УДК 004.93'1

Семантический Анализ Математических Документов в PDF формате

И. А. Меньшиков

Московский государственный университет

Число публикаций по любой тематике, доступных в электронной форме в сети Веб, неуклонно растет. Поэтому проблема поиска новой информации среди похожих документов становится более актуальной. В настоящее время существуют развитые поисковые машины *(Google, Yandex и т.д.),* эксплуатируются базы знаний *(Google Scholar),* представлены системы управления персональными документами *(Mendeley, Zotero)* и социальные сети научных публикаций *(ResearherGate)*. Однако в научном сообществе по-прежнему существует спрос на более универсальное интеллектуальное управление знаниями, представленными в электронных документах, для отдельных пользователей и групп лиц. Задача состоит в том, что необходимо автоматически извлекать знания из документов и интегрировать их в пользовательскую онтологию.

Извлечение знаний из произвольного текстового документа достаточно трудно. Однако документы тематики STEM *(science, technology, engineering, mathematics)* содержат формальные знания [1]. К таким знаниям относятся, в частности, математические формулы и выражения логики. Благодаря тому, что формальные знания часто имеют вполне определенное представление, есть возможность проанализировать их синтаксически и семантически, тем самым получив отдельные понятия и связи между ними, что необходимо для интегрирования информации в онтологию. Исследуемый документ предлагается трансформировать семантически наполненный документ формата OMDoc, поскольку помимо математических формул он поддерживает также специальные элементы математического текста: теоремы, определения и т.д., а для визуального представления документа существует транслятор OMDoc в HTML+MathML.

Неформальные и формальные знания в документе часто перемешаны, что затрудняет извлечение последних [1]. В подавляющем большинстве случаев для анализа доступны только PDF файлы. Поэтому необходимо прежде всего восстановить структуру документа до уровня иерархического представления текста (наподобие формата MS Word).

Для идентификации работы в базе знаний требуется знать её название и авторов. Дополнительные сведения, которые можно извлечь из документа, используются для автоматического разрешения неоднозначностей авторов. Явные связи с другими документами содержатся в библиографии. Наиболее популярной библиотекой разбора библиографии является ParsCit. Данная библиотека разбирает отдельные библиографические ссылки с помощью метода условных случайных полей. Для поиска библиографии в документе не используется знание особенностей изображения текста в PDF. Отмечается, что учет шрифтового оформления и позиции элементов на странице позволяет улучшить распознавание эвристик по сравнению с методами машинного обучения, не использующих данные свойства [2].

В работе предлагается до поиска библиографического раздела найти в тексте регулярные шаблоны и среди них классифицировать библиографические ссылки.

Среди регулярных шаблонов встретятся внутренние ссылки, например, номера рисунков, таблиц и т.д., а также индексы формул. Ссылки на формулы сохраняются для их дальнейшего рабора. Библиографические ссылки далее используются для определения библиографического списка и сегментирования списка на отдельные записи. В случае неудовлетворительного извлечения библиографии выполняется запрос к внешним базам знаний, таким как Google Scholar.

В предыдущей работе [3] разбирались вопросы извлечения математических формул из PDF документа. Синтаксический анализ выделенной математической формулы заключается в линейном представлении на первой стадии и последующий грамматический разбор на второй стадии. Данный подход был выбран для перспективной поддержки разбора математических формул непосредственно из TeX-документов. Анализатор пространственных отношений аналогичен синтаксическим анализаторам 2D грамматик. Применение продукционного правила сводится к использованию классификатора пространственных отношений между близкими элементами. В отличие от грамматик для рукописных текстов, при разборе формулы в PDF можно использовать разбор структуры сверху-вниз с помощью метода нарезки проекций. После нисходящего разбора пространственные отношения ищутся в пределах включающего деления. В случае если отношения не находятся, выбирается область большего размера. Таким образом снижается время работы алгоритма по сравнению с n^3 для алгоритма Коко-Янгера-Касами. При этом увеличивается точность разбора по сравнению с чисто нисходящим алгоритмом.

Формула в линейном представлении формата а-ля TeX разбирается с помощью анализатора, построенного на грамматике в форме EBNF. Грамматика поддерживает математические элементы, указанные в регламенте [4], за исключением матриц и других специальных конструкций. Разрешение неоднозначностей разбора происходит с помощью выбора наиболее приоритетного правила и явного указания ассоциативности операторов

Реализация грамматики на данный момент написана с использованием библиотеки ANTLR4. Исключение неоднозначностей грамматического разбора не исключает содержащихся ошибок, поэтому предусматривается механизм перестроения дерева разбора в случае, если при разборе было выбрано неверное правило. Решение о правильности выбранных правил принимается при заполнении таблицы идентификаторов. Данная таблица заполняется сведениями об использовании идентификаторов в документе в пределах их областей видимости. Полагается, что значение идентификаторов в формулах согласованно в едином контексте. Иерархия документа учитывается при формировании областей видимости.

При анализе текстовой информации предполагается фильтровать неформальную информацию с помощью исследования стилистика фрагментов текста, статистики использования математических понятий и т.д. С помощью адаптивных шаблонов в оставшемся тексте ищутся конструкции логики первого порядка или выше.

Литература

1. *Kohlhase A., Kohlhase M.*. Towards a Flexible Notion of Document Context //2011.
2. *Beel J. et al*. Docear's PDF Inspector: Title Extraction from PDF files //In Proceeding of the 13th ACM/IEE-CS Joint Conference on Digital Libraries. 2013.
3. *Меньшиков И.* Распознавание математических формул в PDF документе // 58 научная конференция МФТИ. 2015