УДК 004.89

Формальная и контекстная проверка текстовых документов

Лушев А.А., Половикова О.Н.
Алтайский государственный университет
andreics1993@mail.ru, pon@asu.ru

Аннотация

В данном исследовании предлагается и рассматривается подход для формального и контекстного анализа текстового документа. Вместо самих текстовых документов предлагается использовать их контекст (в виде rdf–схемы). На основе rdf–схем производиться дальнейшее оценивание текстового документа и выдача рекомендаций. Сами рекомендации формируются на основе Sparql–запросов.

На сегодняшний день известно множество исследований в области онтологий и практических примеров построения семантических поисковых систем, но проблема построения систем контекстного поиска является актуальной. В реальной практике разработчики поисковых систем, учитывающие семантику запроса и контекст самих документов, сталкиваются с различными трудностями. Это связано с несколькими причинами:

• Низкая формализуемость методов семантического поиска и процесса подготовки самих ресурсов.

По этой причине отсутствие в должной мере необходимых технологий для автоматизации всех этапов контекстного поиска от подготовки ресурсов к публикации в глобальной сети до интерпретации найденных семантических структур для пользователей.

• Узкая специализация таких систем.

Семантические поисковые системы, как правило, реализованы для конкретной предметной области, в этой области и работают. Для использования разработанных модулей и программных средств для другой предметной области необходим ресурсоёмкий процесс их адаптации.

• Привлечение дополнительных средств и ресурсов для предварительной обработки базы документов.

Чтобы документ можно было обрабатывать поисковой системой, нужно сформировать и сохранить для него некоторый семантический скелет.

• Ограничения на типы и форматы ресурсов.

Если рассматривать не только текстовые документы (например, графические ресурсы), то возникают сложности различного рада связанные с выделением семантики и её обработкой.

Но, несмотря на трудности проектирования и построения систем контекстного поиска данное направление является перспективным и непрерывно развивается. Предметные области использования систем контекстного поиска разнообразны. Преподаватели учебных заведений и учителя школ сталкиваются с большим объёмом работ по проверке текстов научных исследований. На сегодняшний день в достаточном количестве нет автоматизированных систем оценивания содержания научных работ, полностью отвечающих потребностям проверяющих. Системы, оценивающие смысловое содержание текстовых работ, являются востребованными.

В рамках проведенного исследования был разработан и программно реализован подход для предварительного заключения–оценки текстовых документов и выдачи рекомендаций.

Предлагаемый подход позволяет в полуавтоматическом режиме выполнить проверку текста исследования на соответствие общепринятым стандартам оформления курсовых или выпускных работ, а также ознакомиться с решением поставленных задач в данной работе, не читая ее полностью, что помогает значительно сэкономить время на проверку и оценивание работы.

Суть подхода заключается в следующем. Вместо самих текстовых документов предлагается использовать их контекст (структурированный документ, представленный в виде rdf–схемы). На основе этих rdf–схем будет производиться дальнейшее оценивание текстового документа и выдача рекомендаций. Сами рекомендации выдаются на основе SPARQL –запросов.

Анализ методов и подходов для оценивания работы с учётом семантики контекста показал необходимость создания rdf–схем, которые будут использоваться в процессе работы вместо самих документов. Анализ языка SPARQL для обработки rdf–схем показал возможность его использования для выполнения запросов к содержанию документа.

Рекомендации будут формироваться на основе поэтапного выполнения следующих шагов:

1) происходит автоматическая проверка, на соответствие текста формальным требованиям (Присутствуют ли в данной работе цель, задачи, выводы и другие обязательные пункты).

Для выполнения этого этапа созданы запросы-шаблоны, которые должны выполняться на любом текстовом документе. Если на этом шаге система выдает в качестве ответа пустые запросы, то данный текст не соответствует стандартам, и преподаватель далее её может не оценивать. Если же формальная оценка выдала не пустой результат, то преподаватель (пользователь системы) может увидеть перед собой структурированный текст, отображающий задачи, цель и т.д.

2) выполняется дальнейшая проверка исходного текстового документа отформатированного с учётов поставленных задач.

Такой акцент в форматировании даёт преподавателю (пользователю) наглядное представление того, как решаются поставленные задачи в текстовой работе.На этом шаге, используя сформированный файл в формате html, можно быстро ознакомиться с содержанием решения каждой задачи отдельно. Во втором файле в формате txt каждый блок текста, отвечающий за решение одной из поставленных задач, высвечивается определенным цветом.

Для реализации данной подхода каждую формальную задачу разбили на слова, затем от каждого слова взяли основу, используя для этого алгоритм Портера. Для выдачи содержательной части каждой задачи дополнительно формируется массив рейтинга абзацев, для этого вычислялось вхождение каждой основы в тексте работы. Если вхождение произошло, то в массив рейтинга данному абзацу прибавлялась единица, иначе ничего не происходило, и так для всех основ текущей задачи. Таким образом, для каждой задачи создается свой рейтинг абзацев. Абзац с наибольшим рейтингом записывается в html документ, сформированный в виде шаблона заранее, и также выделяется определенным цветом в исходном текстовом документе. Данная процедура выполняется для каждой формальной задачи.

Таким образом, используя предлагаемый подход и программный модуль можно в автоматическом режиме выполнить проверку текстовой работы на соответствие формальным требованиям. Формируемые дополнительные файлы помогают ознакомиться с содержательной частью исследования.

Список литературы

1. Басипов А.А., Демич О.В. Семантический поиск: проблемы и технологии // Вестн. Астрахан. гос. техн. ун-та. Сер.: управление, вычисл. техн. информ. — 2012. — № 1. — С. 104–111.

2. Половикова О.Н. Анализ способов формализаций документов для выполнения семантического поиска // Известия Алтайского государственного университета. — 2012. — № 1(73). — С. 101–103.

3. Sparql – язык запросов к RDF. — URL: http://www.semanticweb.narod.ru/3.html (дата обращения: 27.10.2016).