Автоматизированное извлечение знаний из научных текстов

Федотов А.М., Барахнин В.Б., Жижимов О.Л., Колчанов Н.А., Федотова О.А.
Институт вычислительных технологий СО РАН, Новосибирский госуниверситет, Институт цитологии и генетики СО РАН, Государственная публичная научная библиотека СО РАН

Использование информационных ресурсов в научно-исследовательском процессе выдвигает необходимость быстрого извлечения "фактов" или "знаний", содержащихся в этом ресурсе.

В связи с этим в данном докладе предлагаются подходы к полуавтоматическому извлечению метаданных из текста (персоны, ключевые слова, оглавление, ссылки) и извлечения фактов (научных результатов в соответствии с онтологией, понятиями) с обеспечением указателей на соответствующие разделы документа, а также средства работы с библиографическими ссылками.

Автоматизация процесса извлечения метаданных и фактов из электронных документов может быть относительно просто реализовано в случае, когда обрабатывается коллекция более или менее однотипных документов, например, статьи с сайта некоторого журнала.

Для извлечения из документов метаданных обычно используются сведения об их разметке в соответствии с правилами формата электронного представления, используемого для хранения документов данной коллекции. Проанализировав особенности представления документов, характерные для обрабатываемой коллекции, пользователь создает соответствующий шаблон, в котором содержатся образцы разметки, применяемой для выделения тех или иных полей метаданных. Таким образом, данный алгоритм основан на анализе *синтаксического* уровня представления информации, при этом речь идет не о синтаксисе естественного языка, на котором написан документ, а о синтаксисе формата электронного представления.

Задача извлечения фактов является гораздо более сложной, поскольку требует проведения анализа на *семантическом* уровне с учетом синтаксической структуры естественного языка документа. Применительно к русскоязычным документам наблюдаются дополнительные трудности, связанные с использованием в русских научных текстах всей широты синонимических рядов общеупотребительных слов, а также свободный порядок слов в предложениях.

Сравнительно простые алгоритмы могут быть использованы для извлечения фактов из коллекций **авторефератов**, поскольку данный тип научных документов имеет более или менее фиксированную структуру, предусматривает четкое выделение основных фактов, а также использование при их изложении достаточно ограниченного количества лексических конструкций.