УДК 004

К.А. Лаврентьев,
ассистент кафедры информационных систем и технологий
Хабаровской государственной академии экономики и права

ОБЗОР МЕТОДОВ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ФОРМ ПЕРВИЧНЫХ ЭКОНОМИЧЕСКИХ ДОКУМЕНТОВ

This article describes methods of economic documents forms parsing. The conclusion about expediency of using any described methods for the universalization of the software for companies' automation workflow.

Keywords: *informational period of development, information technology, electronic document workflow, business processes, Internet services.*

В данной статье автором будут описаны методы синтаксического анализа форм экономических документов и сделан вывод о целесообразности использования какого-либо из описанных методов для универсализации программного обеспечения автоматизации документооборота компаний. XXI в. называют временем, когда человечество вошло в информационный период развития. Сейчас сложно представить себе вид человеческой деятельности, который возможен без использования информационных технологий и автоматизации каких-либо операций. Особенно широко информационные технологии используются в области автоматизации различных бизнес-процессов в корпоративном секторе экономики, связанных с электронным документооборотом. Основная трудность автоматизации документооборота бизнес-процессов компаний заключается в том, что у каждой компании существует множество форм экономических документов и эти формы документов могут различаться. Такие различия не позволяют предложить корпоративному сектору единое универсальное средство автоматизации документооборота. Конечно, существуют различные универсальные решения – продукты фирмы «1С», интернет-сервисы («Мегаплан»), но для полного соответствия документообороту компании данные программы требуют доработки.

Для работы с электронным документооборотом необходимо провести анализ текста. *Анализ текста* – процесс получения высококачественной информации из текста на естественном языке. Как правило, для этого применяется статистическое

обучение на основе шаблонов: входной текст разделяется с помощью шаблонов, затем производится обработка полученных данных. Таким образом, процесс получения информации для построения универсальных форм экономических документов является интеллектуальным, то есть задействуется такая область информатики, как интеллектуальный анализ данных. *Интеллектуальный анализ данных (или Data Mining)* – это собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности (термин введён Г. Пятецким-Шапиро в 1989 г.). Таким образом, при анализе текста нельзя быть уверенным, что в нём будут необходимые данные, это можно только предположить с определённой долей вероятности. В этом и заключается основной вопрос целесообразности использования интеллектуального анализа в системах корпоративной автоматизации. Необходимо понять, целесообразно ли использовать Data Mining при корпоративной автоматизации. Для ответа на этот вопрос необходимо рассмотреть процесс анализа текста и его сложность, а соответственно трудозатраты на программирование этого процесса в системах автоматизации. Перед тем как анализировать текст, необходимо провести ряд подготовительных операций:

– удалить стоп-слова (стоп-слова – это вспомогательные слова, которые несут мало информации о содержании документа («так как», «кроме того»));

– провести морфологический поиск (стэмминг), то есть привести каждое слово к его нормальной форме;

– привести регистр текста к одному виду (обычно к нижнему).

После того как текст подготовлен к анализу, применяют методы Data Mining. Среди этапов анализа текста в Data Mining необходимо отметить:

– классификацию-определение для каждого документа одной или нескольких заранее заданных категорий, к которым он может относиться;

– кластеризацию (автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества);

– автоматическое аннотирование, позволяющее сократить текст, сохраняя его смысл;

– извлечение ключевых понятий (идентификация фактов и отношений в тексте).

Первые два этапа анализа текста относятся к классическому интеллектуальному анализу, при их реализации используются такие методы, как нейронные сети и машинное обучение. Как известно, эти методы построены на математическом анализе, имеют вероятностную природу

(то есть дают неточный результат и долю вероятности вхождения исследуемого объекта в заданное множество) и довольно высокую степень сложности реализации. Однако среди достоинств данных методов можно отметить малую степень ошибки прогнозирования и, как следствие, высокую эффективность работы.

Особенно интересен четвёртый этап анализа текста – извлечение ключевых понятий. Для реализации этого этапа применяют два метода:

– определение частых наборов слов и объединение их в ключевые понятия;

– извлечение ключевых понятий с помощью шаблонов.

Первый метод использует для своей работы базу знаний и классификацию, что говорит опять же о высокой сложности реализации. Извлечение ключевых понятий с помощью шаблонов может помочь отнести исследуемый текст к той или иной категории без применения методов классификации и кластеризации, что позволяет сократить сложность реализации подсистемы прогнозирования при разработке системы автоматизации.

Однако этот метод имеет высокую степень ошибки прогнозирования и даёт не очень точные результаты.

Таким образом, для определения целесообразности применения интеллектуального анализа для получения знаний из форм первичных документов в системах автоматизации предприятий необходимо решить степень важности каждого критерия – простота реализации метода (соответственно малые затраты на разработку и маленькая точность получаемых результатов) или же высокая точность получаемых результатов (но при этом большие затраты на разработку системы).

Литература

1. Яхъяева Г. Э. Нечёткие множества и нейронные сети : учеб. пособие. М. : Бином, 2008. С. 152 – 162.

2. Новейшие методы обработки изображений / под ред. А. А. Потапова. М. : Физматлит, 2008. – 496 с.

3. http://ru.wikipedia.org/