StructuralGeomorphologic Interpretation of the Territory ofThe Republic of Khakassia According to Remote Sensing

Ksenia VladislavovnaShatrova,Assistant, Siberian Federal University

Tat'ana AlexandrovnaJankovskaja, Candidate of Physico-Mathematical Sciences, Associate Professor Siberian Federal University

YuriAnatol'evich Maglinets, Candidate of Technical Sciences, Professor Siberian Federal University

The paper is considered the technique structural geomorphic interpretations data the Earth remote sensing of territorial the Republic of Khakassia,conducteda structural analysis ofthe elementarylevelof the landscape structure.

Keywords:landscape, Earth remote sensing data, geomorphologicalinterpretation ofthe territory.

УДК 004.912

ПРИМЕНЕНИЕ ОНТОЛОГИИ ЖАНРОВОЙ СТРУКТУРЫ ДЕЛОВОГО ДОКУМЕНТА В ПРОЦЕССЕ ВЫДЕЛЕНИЯ РЕКВИЗИТОВ ДЕЛОВОГО ТЕКСТА

***Екатерина Михайловна Гриценко**,к.т.н., доцент,*
Тел.: 8 923 3541985,e-mail: mmlab@bk.ru,
***Владимир Васильевич Гуменюк**,аспирант*
Тел.: +7 923 2862076, e-mail: cruelled@gmail.com
ФГБОУ «Сибирский государственный технологический университет»
http://www.sibstu.kts.ru

В статье рассмотрено применение метода анализа реквизитов делового документа, основанного на использовании онтологии жанрового уровня структуры документа. Применение метода рассматривается на основе обработки документа «Требование», рассылаемого Межрайонной ИФНС России №22 по Красноярскому краю.

Ключевые слова: онтологии, реквизит, деловой документ, жанровая структура документа.

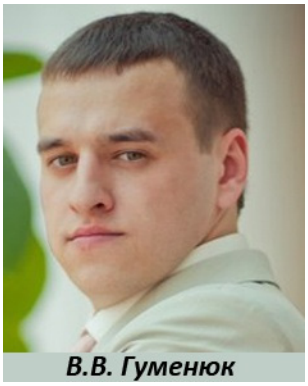
В современном, деловом мире объёмы переписки неуклонно растут. Многие деловые документы (ДД), в процессе согласования, несколько раз пересылаются между субъектами переписки.

Использование информационных технологий и средств телекоммуникаций позволяет обмениваться огромными объёмами информации практически мгновенно, однако обработка её получателем связанна с рядом трудностей.

Так обработка полученной информации может быть возложена на персонал компании получателя, что обеспечит гибкость, и точность обработки, однако связано со значительными трудовыми и временными затратами.


Е.М. Гриценко

Системы автоматической обработки текста, как правило, пытаются определить, что именно написано в тексте. Такие системы весьма чувствительны к базе знаний, с которой они работают, а также чувствительны к вычислительным ресурсам, обеспечивающим их скорость работы. Результат работы таких систем затрудняет их использование для решения задач требующих стабильности и точности. Внедрение систем электронного доку-


В.В. Гуменюк

ментооборота позволяет решить проблему обработки больших объёмов полученной информации в сжатые сроки, однако они требуют первоначальных вложений, которые могут быть неподъёмны для субъектов малого предпринимательства, и главное введения единого стандарта обмена информацией. Многие министерства и ведомства Российской федерации не могут внедрять такие систем в виду, того, что список используемого ими программного обеспечения жёстко ограничен.

В группе систем автоматически обрабатывающих текст особняком стоят системы, использующие различные онтологии для анализа текстов на естественном языке (Е.Я.). Они позволяют легко наращивать базу знаний для своей работы, но требовательны к вычислительным ресурсам. В статье [1], был изложен подход к анализу текстов на Е.Я. основанный на использовании онтологии жанровой структуры ДД для определения реквизитов ДД. Целью данной статьи является приведение результатов использования данного подхода при анализе ДД передаваемых Межрайонной ИФНС России №22 по Красноярскому краю.

Одной из услуг оказываемых клиентам, почтой России, является печать, с последующей доставкой, писем из предоставленного клиентом электронного файла, содержащего текст соответствующих почтовых отправлений. Что требует определение реквизитов получателя и отправителя.

В виду ограниченности трудовых ресурсов ручная обработка не представляется возможной. На данный момент выполнение этих операций осуществляется с помощью программного комплекса Letters Elaboration. Однако формат писем поддерживаемых этим комплексом чётко задан, что ведёт к необходимости периодического внесения изменений.

К обработке принимаются документы в формате «*.doc», «*.rtf», что требует регулярной доработки комплекса при изменении формальной структуры документа. Прийти к единому неизменному формату данных не представляется возможным в виду различных административно технических причин. Так же комплекс не способен провести предварительный анализ документа на соответствие заданному шаблону, что в ряде случаев приводит к не корректному определению атрибутов документа и как следствие не верное определение реквизитов письма. На данный момент диагностирование этих проблем возложено на оператора комплекса, что приводит к дополнительным материально временным затратам.

Для решения этой задачи разрабатывается метод автоматического распознавания атрибутов делового документа. Этот метод должен корректно распознавать атрибуты делового документа, основываясь на жанровой структуре документа. Для её описания удобно использовать подход, использующий онтологии.

Общая схема обработки делового документа

В рамках предлагаемого подхода общая схема обработки делового текста представляется в виде совокупности нескольких этапов:

- получение электронной копии документа;
- проведение сегментации;
- идентификация сегментов на основе онтологии структуры делового документа;
- представление документа в виде совокупности атрибутов;
- распознавание содержания требуемых атрибутов.

Схематично процесс обработки делового документа (ДД) представлен на рисунке 1.

Первый этап заключается в получении электронной формы ДД а, которая может быть получена в результате сканирования ДД с последующим распознаванием, изначальным созданием в текстовом редакторе или любым другим способом. По завершении первого этапа электронная форма ДД переходит на стадию сегментации.

На втором этапе, на стадии сегментации электронная форма ДД ***a***, с помощью лексического анализа и/или другого метода разбивается на сегменты ***b***, признаками которых является графическая изолированность от остального текста.

На третьем - четвёртом этапах обработки, сегментам текста ***b*** ставятся в соответствие экземпляры онтологии ***c***, результатом чего является представление ДД в виде совокупности атрибутов делового текста ***d.***

В зависимости от поставленных задач на последнем этапе происходит обработка необходимых атрибутов ДД, в том числе и распознавание атрибутов ***e.***

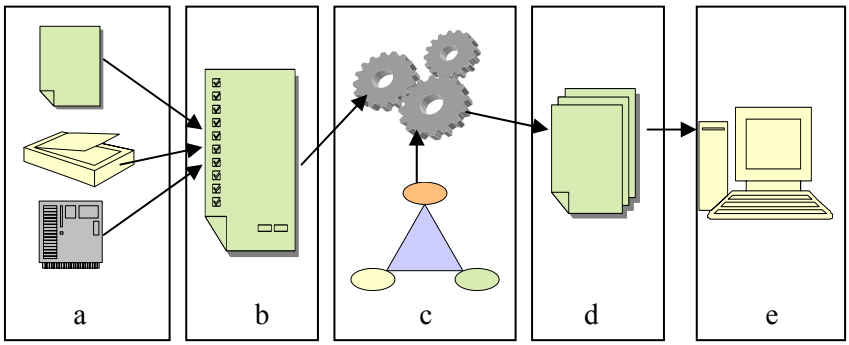


Рис. 1. Общая схема обработки документа
a - электронная форма делового документа; b - сегменты текста; c - онтология делового документа; d - совокупность атрибутов делового текста; e - распознавание атрибутов

Представление знаний

Для функционирования системе необходимы знания о возможных атрибутах делового документа, а так же о взаимоотношениях между ними. Один из возможных подходов, для описания необходимой информации основывается на использовании онтологии.

Учитывая специфику задачи можно сказать что, онтология включает в себя таксономию классов, описывающих различные атрибуты ДД, связанный с классами тезаурус и множество связей, описывающих как классы взаимосвязаны между собой. Каждый класс в онтологии может обладать рядом свойств, а так же ограничений, как на значения свойств, так и на отношения в которых участвует этот класс. В некоторых источниках экземпляры класса выносятся в отдельное подмножество. Однако в рамках данной задачи, под экземпляром класса мы будем понимать потомок класса родителя с заполненными свойствами.

Под областью ДД понимается некоторая устойчивая область ДД содержащая ряд атрибутов ДД объединённых единым назначением. Так на верхнем уровне, ДД состоит из областей «Начало ДД», «Текст ДД», «Завершение ДД», «Приложения ДД». Так же в онтологию вводятся классы, обозначающие все атрибуты ДД описанные в госте [3].

При проектировании онтологии описывающей жанровую структуру делового текста, необходимо ввести ряд отношений, которые будут описывать взаимное расположение атрибутов ДД, приведённых ниже:

- является частью (part_of) – в общепринятом значении;
- является экземпляром (is_a) – в общепринятом значении;
- находится правее – отношение А находится правее Б означает, что в документе атрибут А расположен правее атрибута Б;
- находится ниже - отношение А находится ниже Б означает, что в документе атрибут А расположен ниже атрибута Б.

Конкретное задание отношений, и ограничений для классов онтологии, определяется исходя из госта [3]. Таким образом, спроектированная онтология позволяет описать жанровую структуру любого ДД

Особенности вывода

После сегментации текста происходит процесс идентификация сегментов на основе онтологии структуры делового документа. В рамках которого, сегменту расположенному левее и выше всех ставится в соответствие класс в онтологии на основании свойств класса.

В дальнейшем, исходя из отношений между классами онтологии сегменту С1, расположенному правее стартового сегмента Сс, ставится в соответствие класс онтологии К1, для которого установлено отношение «находится правее» с классом сегмента Кс. Аналогично происходят идентификация на основе остальных отношений в онтологии. Так сегмент С2 ставится в соответствие с классом К2 в силу того, что между классом К1 и классом К2 установлено соответствие «находится ниже». На основе отношения «находится правее» между классами К3 и К2 сегменту С3 ставится в соответствие класс К3.

Необходимо отметить, что каждому классу, которому ставится в соответствие сегмент, соответствует атрибут делового текста.

Таким образом, в результате процесса мы получаем текст в виде набора атрибутов делового текста, что позволяет в дальнейшем анализировать содержание только значимых для конкретной прикладной задачи атрибутов.

Используемая методология

При построении методологии жанровой структуры ДД использовалась методология methontology [2]. В рамках данной статьи, приведены три первых этапа, methontology:

- построение глоссария терминов;
- построение деревьев классификации концептов;
- построение диаграмм бинарных отношений.

Глоссарий терминов

Глоссарий терминов включает все термины (концепты и их экземпляры, атрибуты, действия и т. п.), важные для предметной области, и их естественно-языковые описания.

В таблице 1 приведён фрагмент словаря терминов, использованного при построении онтологии жанровой структуры делового текста.

Таблица 1

Фрагмент словаря терминов

Название класса	Описание класса
Начало документа	Часть электронной версии текста на Е.Я. идентифицирующая документ.
Текст документа	Часть электронной версии текста на Е.Я. несущая основную мысль документа.
Конец документа	Часть эл-й версии текста на Е.Я. однозначно идентифицирующая окончание документа.
Область утверждения и контроля	Область, в которой содержится информация об утверждении документа
Область адресата	Область, в которой записывается информация о том, кому адресован документ
Область заглавия	Область, содержащая заголовок документа
Текстовая область	Область, содержащая основной текст документа
Визирующая область	Область, в которую размещаются подписи заверяющие документ
Область согласования	Область, в которой содержится информация о согласовании документа
Область исполнения	Область, в которой содержится информация о состоянии исполнения документа

Древо классификации концептов

При достижении словарём терминов, существенного объёма, термины организуются в деревья классификации концептов. Деревья классификации концептов описывающих таксономию классов при данном описании предметной области. Таким образом, идентифицируются основные таксономии предметной области. Фрагмент древа классификации концептов использованного, при построении онтологии жанровой структуры делового текста приведён на рисунке 2.

Диаграммы бинарных отношений

Целью создания диаграмм бинарных отношений, является фиксация отношений между концептами одной или разных онтологий. Диаграммы бинарных отношений, фиксирующие отношения между несколькими концептами онтологии жанровой структуры делового текста приведены на рисунках 3-5.

Применение онтологии жанровой структуры делового текста для выделения реквизитов делового документа

В качестве примера проведём анализ документа рассылаемого Межрайонной ИФНС России №22 по Красноярскому краю

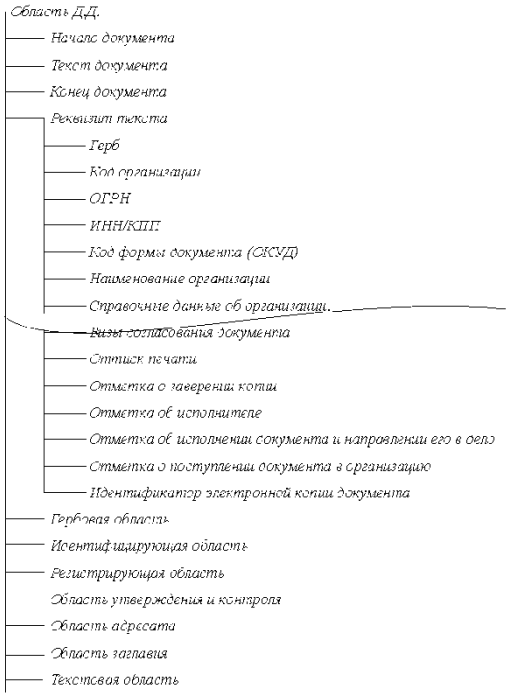


Рис. 2. Фрагмент древа Классификации концептов

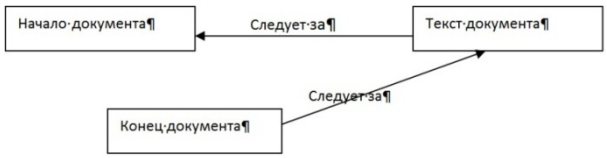


Рис. 3. Диаграмма бинарных отношений между элементами страктами «Деловой документ»

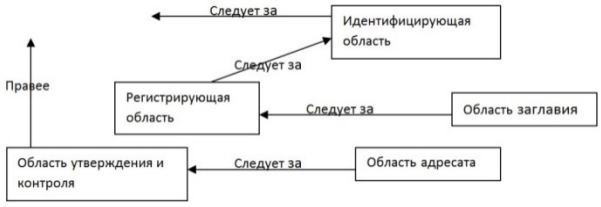


Рис. 4. Диаграмма бинарных отношений между элементами стракта «Начало документа»

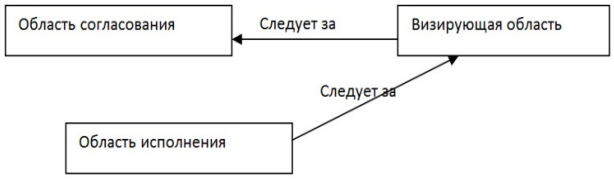


Рис. 5. Диаграмма бинарных отношений между элементами стракта «Конец документа»

Так, исходя из графического расположения текста, документ можно разделить на 3 глобальные блока, а именно «Начало документа», «Текст документа», «Конец документа», которые можно видеть на рисунках 6-8 соответственно.

В блоке «Начало документа» (рисунок6), можно выделить 3 графически изолированные друг от друга зоны (***a***,***b***,***c***). Исходя из диаграммы бинарных отношений представленных на рис. 3 зона (***a*** рисунок6) может являться гербовой областью, однако отсутствие герба опровергает это предположение. «Идентифицирующая область» связана с гербовой областью отношением «следует за», область (***a*** рисунок 6) вероятно соответствует атрибуту «Идентифицирующая область», наличие в тексте области, индекса подтверждает это предположение.

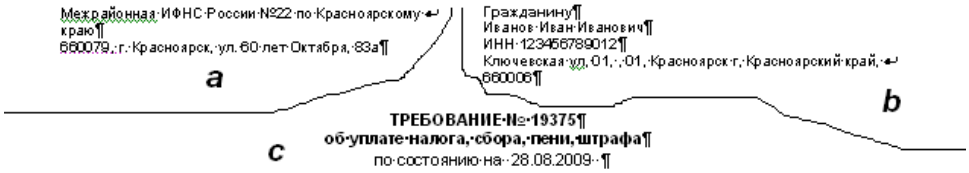


Рис. 6. Экземпляр концепта «Начало документа», с отмеченными составляющими его экземплярами концептов.

Область (***b*** рисунок 6) находится правее области (***a*** рисунок 6) и соответственно диаграмме на рис. 3 может являться экземпляром концептов «Область утверждения и контроля» и «Область адресата». Наличие в тексте обращения «Гражданину», а так же индекса опровергает предположение о том, что это область является экземпляром концепта «Область утверждения и контроля», соответственно область (***b*** рисунок6) является экземпляром концепта «Область адресата».

Область (***c*** рисунок 6) следует за областью (***a*** рисунок 6), следовательно, она может быть экземпляром концепта «Регистрирующая область», однако выравнивание текста в ней по центру страницы делает это предположение не верным. Следовательно,

область (***c*** рисунок 6) является экземпляром концепта «Область заглавия», так как он связан отношением «следует за» с концептом «Регистрирующая область».

Представленный на рисунке 7 фрагмент области «текст документа» содержит единственный атрибут «Текст».

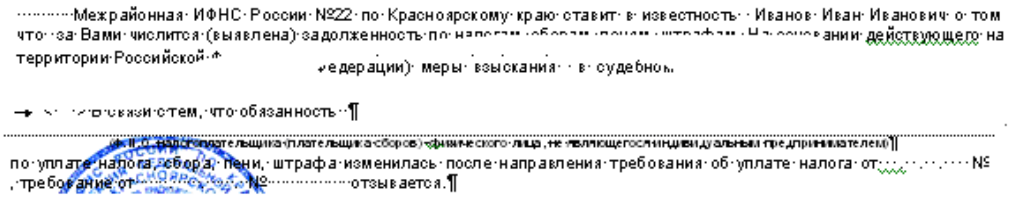


Рис. 7. Экземпляр концептов «Текст документа», и «Текст»

На рисунке 8 представлен экземпляр концепта «Конец документа» в котором можно выделить две области отмеченные как (***a, b***). Исходя из диаграммы бинарных отношений представленной на рис. 4, область (***a*** рис. 7) может быть экземпляром концепта «Область согласования» однако отсутствие слова «Согласованно» опровергает это предположение, следовательно (***a*** рис. 7) является экземпляром концепта «Визирующая область» что подтверждает наличие печати.

Область (***b*** рис. 7) следует за областью (***a*** рис. 7) и, исходя из диаграммы бинарных отношений представленной на рис. 4 является экземпляром концепта «Область исполнения», связанного с концептом «Визирующая область» отношением «Следует за».

Отмеченная область (***a*** рисунок 8) соответствует экземпляру концепта «Визирующая область», область (***b*** рисунок 8) соответствует экземпляру концепта «Область исполнения».

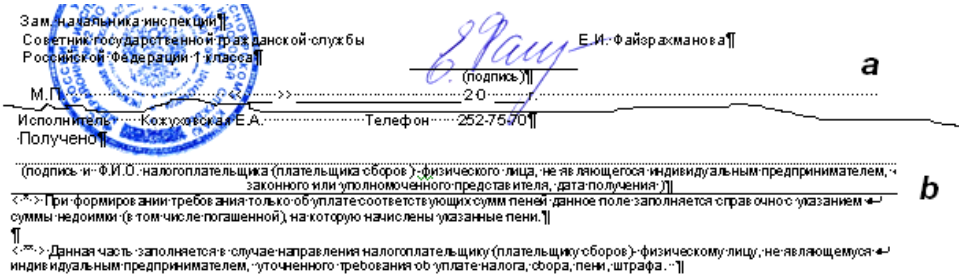


Рис. 8. Экземпляр концепта «Конец документа»

Заключение. В данной статье рассмотрен пример применения подхода к анализу текста делового документа, основанный на использовании онтологии жанровой структуры делового текста. В статье приведены фрагменты глоссария терминов, древа классификации концептов и диаграмм бинарных отношений, использованных в построенной онтологии. Так же в статье приведено описание логического вывода, позволяющего разбить деловой документ, на независимые области, содержащие различные реквизиты делового текст.

Приведённый эксперимент показал, что при использовании данного подхода, удаётся, не проводя анализа всего текста документа, выделить области с требуемой информацией, для дальнейшего детального анализа.

Application of ontology of genre structure of the business document in the course of allocation of requisites of the business text

Ekaterina Mikhailovna Gritsenko, candidate of technical Sciences, associate Professor
Siberian state technological University
Vladimir Vasilyevich Gumenyuk

The paper considers the application of the method of analysis of the business details of the document, based on the use-consistent ontology genre level document structure. The application of the method is considered on the basis of document processing «requirement» rassylaemo the Inter district IFTS Russian number 22 in the Krasnoyarsk Territory.

Keywords: ontology, props, business document, genre structure of the document.