УДК 004.912: 004.632

В.В. ЛИТВИНОВ[*], О.П. МОЙСЕЕНКО[*]

АВТОМАТИЗИРОВАННАЯ СИСТЕМА ОБРАБОТКИ ДИНАМИЧЕСКИХ КОЛЛЕКЦИЙ РАЗНОЯЗЫЧНЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ ПО МОРСКОМУ И РЕЧНОМУ ДЕЛУ

[*]Черниговский национальный технологический университет, Чернигов, Украина

Анотація. *Представлена система автоматизованої обробки великих об'ємів динамічної текстової інформації. Система виконує функції пошуку, класифікації, рубрикації та кластеризації текстових документів за запитами користувача.*
Ключові слова: *класифікація, рубрикація, кластеризація, обробка текстових документів.*

Аннотация. *Представлена система автоматизированной обработки больших объемов динамической текстовой информации. Система выполняет функции поиска, классификации, рубрикации и кластеризации текстовых документов по запросам пользователя.*
Ключевые слова: *классификация, рубрикация, кластеризация, обработка текстовых документов.*

Abstract. *The system of automated processing of large volumes of dynamic, text information was represented. The system performs search functions, classification, categorization and clusterization of text documents at user requests.*
Keywords: *classification, categorization, clusterization, processing of text documents.*

1. Введение

Морское и речное дело – совокупность знаний, относящихся к судоходству: морское и речное право, морская и речная практика, кораблевождение, портовые обычаи, исторические сведения.

Документ – електронный текстовый документ на естественном языке, содержащий структурированную информацию, относящуюся к конкретной предметной области и представленный в одном из распространенных форматов хранения текстовых данных.

Все текстовые документы можно поделить на три документопотока [1]:

– входящий документ – документ, поступивший в учреждение;

– исходящий документ – официальный документ, отправляемый из учреждения;

– внутренний документ – официальный документ, не выходящий за пределы подготовившей его организации.

Каждый из документопотоков имеет свои особенности обработки в зависимости от формы организации работы с документами (централизованный, децентрализованный, смешанный).

Разрабатываемая автоматизированная система (АС), в первую очередь, направлена на обработку первого вида документопотока. Классическая схема ручной обработки входящих документов применима и при автоматизации процесса.

К входящим документам относятся те, которые поступают из других организаций: вышестоящих, подчиненных, общественных, муниципальных, негосударственных, от юридических и физических лиц. Например, законы, указы, постановления, решения, указания, инструкции, распоряжения, поручения, приказы, доверенности, договоры, письма по электронной почте с прикрепленными документами (предложения, сопроводительные, гарантийные, рекламные, информационные и др.), отчеты о хозяйственной, финансовой, управленческой деятельности, докладные записки, акты.

Обработка входящих документов включает:

ISSN 1028-9763. Математичні машини і системи, 2014, № 2

– прием и первичную обработку электронных документов (до автоматизации осуществляется службой делопроизводства,секретарем или специальными подразделениями – экспедициями);

– предварительное рассмотрение и распределение документов (отнесение документов к категории, требующей специального рассмотрения руководством организации или структурных подразделений для того, чтобы освободить руководителя от рассмотрения второстепенных вопросов, ускорить движение документов);

– регистрацию (учетный порядковый номер документа и дата поступления, при необходимости, часы и минуты. Внесение сведений о документе в електронный журнал);

– рассмотрение документов, принятие решения по информации, содержащейся в документе. Возможность правки и удаления содержимого. Изменение приоритетов доступа;

– передачу на исполнение (документы должны исполняться в срок. Входящие документы обязательно передаются исполнителю в день его получения и регистрации или в первый рабочий день при поступлении документов в нерабочее время);

– хранение документов (хранение архивов документов, индексов документов, адресов удаленного доступа к документам).

2. Постановка задачи

Проектирование и разработка системы автоматизированной обработки динамичесих коллекций, разноязычных тектовых документов по морскому и речному делу. Поиск и предоставление документов происходит среди локальных, переносных и удаленных известных источников, а также в глобальной сети Интернет. Результатом работы системы являются тематические кластеры текстовых документов, построенные согласно запросам пользователей (рис. 1). Документы внутри одного кластера должны быть максимально схожи между собой. Общий набор кластеризуемых документов не может быть заранее определен, так как на вход системы непрерывно поступают новые документы.

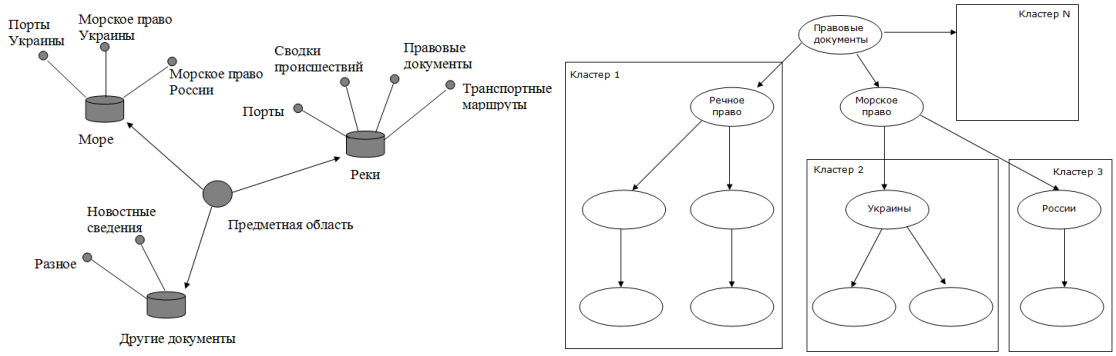

Рис. 1. Примерные схемы кластеризации документов по морскому и речному делу

3. Решение задачи

Значительная часть информации по морскому и речному делу подается на английском – общепринятом языке межнационального общения. Но и на других языках мира содержится очень большая и важная часть информации. Поэтому чрезвычайную значимость и ценность приобретает возможность межъязыковой коммуникации.

На сегодня в мире имеется множество инструментов, позволяющих пользователям понимать получаемую информацию и представлять свои электронные документы на большом числе естественных языков. Это программы проверки орфографии и грамматики, программы автоматического перевода, системы диктовки, пакеты информационного поиска.

Существующие подходы в сфере автоматического перевода текстовых корпусов еще далеки от идеального: перевод имен собственных, неправильная структура предложения, отсутствие грамматических связей и т.д. Неоспоримым преимуществом автоматического перевода являются быстрота и сравнительная, относительно ручного перевода, дешевизна обработки текста. Однако риск возникновения грубых тематических ошибок повышается в случае узкоспециализированного перевода, когда требуется высококвалифицированный переводчик и отменный специалист в конкретной области в одном лице.

В разрабатываемой АС обработки разноязычных тектовых документов уклон сделан в сторону перевода не содержимого документов, а сформированных пользователями запросов. То есть сформированный на понятном для пользователя языке запрос будет интерпретирован с учетом тематической (семантической) составной на другие, доступные системе языки. После чего происходит поиск необходимой информации среди документов на других языках. Результаты поиска представляются пользователям в виде документа(ов) на языке оригинала и при необходимости могут быть переведены сторонними программными продуктами.

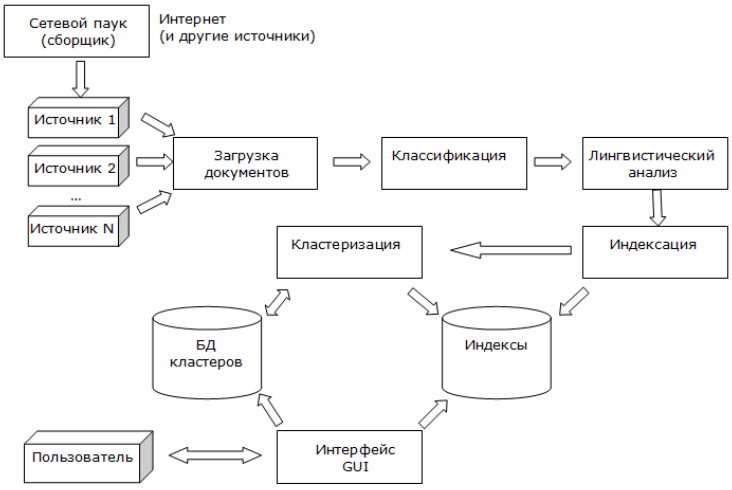

Рис. 2. Структура разрабатываемой системы

Разрабатываемая система имеет клиент-серверную архитектуру (тонкий клиент) и состоит из независимых программных модулей (рис. 2), что в значительной мере повышает отказоустойчивость и надежность такой системы. Принципы работы программной реализации системы описаны ниже.

Подсистема поиска и загрузки – сетевой паук, получает на входе адреса веб-ресурсов (источники текстовых документов). Загружает веб-страницу и сканирует ее в поисках других гиперссылок (если это разрешено). Обнаруживаемые гиперссылки фильтруются и проверяется, были ли они посещены в этом сеансе поиска. Веб-страница проходит обработку для поиска паттерна или индексирования (с целью ускорения последующего доступа к источнику). Происходит загрузка веб-ресурсов по найденным и отобранным гиперссылкам, контролируя трафик и глубину погружения по гиперссылкам.

Среди содержимого сайта загружаются и индексируются архивы и электронные текстовые файлы с расширениями *.zip, *.rar, *.txt, *.doc, *.pdf (список можно расширить). Каждый из форматов хранения данных является программным контейнером и требует реализации в системе отдельных механизмов извлечения этих данных.

После чего происходит создание индексной базы и индексирование содержимого web-источников. Завершение формирования индексной базы позволяет выполнять поиск на сервере даже не имея фактического доступа в сеть.

Предварительная обработка электронных текстов

1. Получение текста документа и последующая его первичная обработка (идентификация формата, языка, кодировки документа, при необходимости приведение к единой кодировке utf8, очистка текста от элементов оформления и графики, разбиение на составные части).

2. Лингвистический анализ (графематический, морфологический и постморфологический анализ, выделение словосочетаний).

3. Формирование векторного (матричного) представления текстов.

Классификация составляющих текстовых документов по темам и подтемам [2]

1. Последующая обработка текстов (отображение словаря признаков документа в пространство признаков классификатора; оценка адекватности и возможности классификации текста с помощью данного классификатора).

2. Классификация текста и выделение значимых фрагментов в нем (выделение кодов рубрик с помощью регулярных выражений; применение логических правил, построенных экспертами, и статистических решающих правил; корректировка результатов классификации с учетом иерархической структуры рубрик).

Анализ результатов классификации (рубрицирование) [2]

1. Выявление "почти" дубликатов документов (документы с измененной синтаксической структурой, но с одинаковым смысловым содержанием. Возникают в случае применения синонимайзеров и услуг рерайтеров).

2. Выявление основных тем документов в рубриках.

3. Упорядочивание документов по их релевантности рубрике.

4. Формирование решающих правил и оценивание качества обучения.

5. Обучение происходит в процессе обнаружения специфических для каждой из рубрик терминов и формирования для каждого термина численной меры значимости, а также порогового значения поискового веса.

6. Формирование обучающих и тестовых множеств для рубрик (построение разбиения обучающего массива на блоки; анализ взаимосязей и пересечений отдельных рубрик; формирование множеств отрицательных и положительных примеров).

7. Оценивание параметров базовых моделей рубрик (вычисление весов признаков; снижение размерности; оценка параметров моделей; формирование решающих правил; оценка качества обучения).

8. Построение комбинированных решающих правил для отдельных рубрик и классификатора в целом.

9. Формирование отчета о результатах обучения (описание решающих правил, описание терминологии рубрик, рекомендации по корректировке примеров документов, описание взаимосвязей рубрик).

Корректировка обучающего массива и настройка правил классификации [3]

1. Обучающая выборка (training sample) – выборка, по которой производится настройка (оптимизация параметров) системы.

2. Корректировка обучающих примеров для рубрик путем анализа добавленных и пропущенных документов в рубриках значимых фрагментов, взаимосвязей рубрик.

3. Настройка правил классификации для отдельных рубрик (явное задание предпочтительных статистических моделей, задание необходимых, достаточных и исключающих логических правил на специальном языке).

Индексирование [4]

1. Среди множества документов, количество и размер которых могут быть очень большими, отбираются только те из них, которые отвечают какому-либо условию, например, содержат ту или иную фразу.

2. Работа с индексами происходит при помощи свободной программной библиотеки Lucene, которая может сохранять / извлекать в индексе оригинальное (неизменное) значение. Нести дополнительную информацию о найденном документе. Позволяет задейство-

вать механизмы анализа содержимого данного документа на этапе создания индекса (выделение слов из набора букв и пробелов между ними). Хранит дополнительную информацию о позициях тех или иных слов в теле документа, что значительно ускоряет процесс поиска найденных вхождений, чем при последовательном обходе и поиске в каждом из документов.

3. Lucene создает своеобразный мост между индексатором (сетевым пауком) и локальним поиском среди коллекции документов. Содержит общую информацию и используется совместно обоими модулями.

Кластеризация текстовых коллекций

1. Вычислительное определение наличия и состава тематически (содержательно) однородных групп в текстовой коллекции в случае, когда априорное описание групп отсутствует.

2. В результате кластеризации для каждой из найденных тематических групп определяются состав группы (список входящих в группу документов), ключевые слова и аннотация группы, дающие пользователю агрегированную информацию о тематике документов группы.

3. Основные тематические группы по запросу пользователя могут быть дополнительно объединены в более крупные группы, а внутри каждой из групп могут быть выделены более мелкие тематические подгруппы (проведена вторичная кластеризация).

Доступ к информации

1. Разбор запроса к поисковой системе от пользователя и/или администратора.

2. Использование возможностей поисковой системы для предоставления ответа пользователю (полнотекстовый поиск, ограничения найденного набора определенным доменом – диапазон дат, рубрик, авторов и др.).

3. Внесение изменений к весу, ранжирование того или иного элемента документа, отдельного термина и последовательности лексем.

4. Просмотр последовательности принятия решения об отнесении документа к определенной теме или поисковому домену.

5. Сохранение и предоставление статистических данных о работе системы.

Прямое общение пользователей (чат)

1. Обмен короткими текстовыми сообщениями (чат) при помощи функционала графического пользовательского интерфейса(GUI) между активными, то есть, находящимися в данный момент в системе, пользователями.

2. Чат является отдельным программным модулем, скрытым за общим пользовательским интерфейсом.

3. Вариации структуры чата:

“head-to-head” есть только один канал, с одной стороны которого сервер, с другой – клиент. Multy-user-структура – один сервер и множество клиентов. Сервер при этом выполняет обработку входящих сообщений, пересылает их по нужным каналам, регистрирует пользователей и показывает всем, сколько пользователей общаются в текущий момент.

Консультация в реальном времени (On-line)

Консультация необходима в случаях, когда пользователь по тем или иным причинам не смог найти интересующую его информацию (неправильная формулировка запроса, нет информации в базе данных), а также в ряде других случаев.

Вариации консультации: текстовая консультация (подобие чата между экспертом и пользователем), телефония (горячие номера call center), интернет-телефония (реализация

аналога программ (Skype, Viber), входящего в состав системы, или использование оригинальных продуктов).

3. Выводы

Предложенная автоматизированная система обработки больших динамичных коллекций разноязычных електронных документов по морскому и речному делу позволит оптимизировать затраты на обработку текстовой информации. Предоставит доступ к важной информации (правовые документы, резолюции, новостные сводки и др.) на всех основных языках мира. Станет универсальным инструментом для организаций и учреждений, работающих в сфере морских и речных дел.

СПИСОК ЛИТЕРАТУРЫ

1. Типовая технология работы с документами [Электронный ресурс]. – Режим доступа: http://www.delcomp.ru/002_9.html.
2. Литвинов В.В. SVM при классификации мультиязычных текстов / В.В. Литвинов, О.П. Мойсеенко // Вісник ЧДТУ. – 2013. – № 4.
3. Пескишева Т.А. Параллельная реализация алгоритма обучения системы текстовой классификации / Т.А. Пескишева, Е.В. Котельников // Вестник УГАТУ. – 2011. – Т. 15, № 4 (44). – С 130 – 136.
4. Открытая программная библиотека Lucene [Электронный ресурс]. – Режим доступа: http://lucene.apache.org/core.
5. Международное морское право: справочник / Под ред. С.Г. Горшкова. – М.: Воениздат, 1985.
6. Weston J. Support Vector Machines for Multi-Class Pattern Recognition [Электронный ресурс] / J. Weston. – Режим доступа: https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es1999-461.pdf.

Стаття надійшла до редакції 20.01.2014