OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.912

ПОИСК И РЕФЕРИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ В МНОГОЯЗЫЧНОЙ СРЕДЕ

Липницкий С.Ф., Мамчич А.А., Степура Л.В.

Объединенный институт проблем информатики НАН Беларуси,
г. Минск, Республика Беларусь
lipn@newman.bas-net.by
lexamam@newman.bas-net.by
stepura@newman.bas-net.by

Рассмотрена технология поиска и обработки научно-технической информации из различных источников (Интернет, локальная сеть, жесткий диск компьютера пользователя) в многоязычной среде. Предложена архитектура информационной системы. Представлены ее основные функции: индексирование текстовых документов, их поиск и автоматическое реферирование.
Ключевые слова: информационный поиск; индексирование; автоматическое реферирование; текст.

Введение

В докладе рассматривается автоматизированная система поиска и реферирования текстовой информации в многоязычной среде, разработанная в ОИПИ НАН Беларуси. Реализованные в системе модели и алгоритмы позволяют активизировать и эффективно использовать информационные ресурсы из различных источников (Интернет, локальная сеть, жесткие диски отдельных компьютеров пользователей).

Предложенный в данной работе подход к поиску и реферированию текстовой информации, в отличие от существующих, основан на использовании тематических и динамических корпусов текстов (совокупностей текстов по конкретной тематике), что обеспечивает адаптацию системы к решаемой задаче и информационным потребностям пользователей, а также независимость программного обеспечения от входных языков. Такие корпусы текстов могут создаваться предварительно под прогнозируемые задачи, а также формироваться оперативно после поступления запроса (динамические корпусы текстов).

1. Архитектура системы

Функциональными компонентами системы поиска и аналитической обработки научно-технической информации являются три подсистемы:

– подсистема индексирования текстовых документов в Интернете, локальной сети и на жестком диске;

– подсистема информационного поиска;

– подсистема реферирования текстовых документов.

В состав подсистемы индексирования входят следующие информационно-программные средства:

– программы выявления информативных слов и предложений. Используются при индексировании текстовых документов и веб-страниц;

– программы индексирования текстовых документов и веб-страниц. Каждому документу (странице) приписывается совокупность поисковых признаков с их весами (числовыми значениями информативности).

Подсистема информационного поиска включает:

– программы поиска текстовых документов и веб-страниц. Результатом поиска являются документы и адреса страниц, упорядоченные по убыванию их информативности;

– программы поиска по содержанию. Реализуют фактографический информационный поиск в полнотекстовых документах;

Подсистема реферирования текстовых документов состоит из программ выявления информативных предложений и синтеза связных рефератов.

Предложенный разработчиками метод поиска и реферирования текстовой информации обеспечивает функционирование системы в многоязычной среде. Адаптация программного

комплекса к новому входному языку не требует доработки и корректировки программ. Необходимо лишь сформировать в базе данных корпус текстов на этом языке. Процедуры создания словарей базы знаний реализуются в автоматизированном режиме. В настоящее время пользователи системы могут работать на русском, белорусском, английском и немецком языках.

Все подсистемы реализованы в виде единого программного приложения, что позволяет наиболее эффективно организовать процесс работы пользователей и администратора информационной системы.

1.1. Корпусы текстов

Задачи создания и использования корпусов текстов решаются в рамках специального раздела языкознания – ко́рпусной лингвистики. Под корпусом текстов понимают совокупность документов, накопленных и размеченных по определенным правилам в зависимости от назначения. В случае отсутствия разметки эти совокупности называют корпусами текстов первого порядка. Различают тематические корпусы текстов (наборы текстов по предметным областям) и полные корпусы текстов, каждый из которых объединяет все тематические корпусы на данном входном языке. Для каждого языка (например, русского, белорусского, английского) создается свой полный корпус текстов.

1.2. Словари базы знаний

1.2.1. *Словари словоформ и парадигм*

В словаре словоформ каждой словоформе поставлены в соответствие:

– частота в полном корпусе текстов;

– частоты во всех тематических корпусах текстов;

– номер (код) парадигмы.

Словоформы и их статистические характеристики хранятся в словаре словоформ. В первоначальном состоянии каждая словоформа словаря образует отдельную парадигму. После объединения некоторых (или всех) словоформ в словоизменительные парадигмы словоформам присваивается номер парадигмы, элементом которой эта словоформа является.

Словарь парадигм служит для поиска всех словоформ парадигмы после нахождения словоформы и ее кода в словаре словоформ. Процедура поиска используется при вычислении информативности слов. Создается и актуализируется словарь парадигм в человеко-машинном режиме с использованием соответствующего инструментария. В первоначальном варианте каждая парадигма словаря парадигм содержит одну-единственную словоформу для каждого кода словоформы. После формирования парадигм коды меняются.

1.2.2. *Словарь синонимичных словоформ*

Словарь состоит из групп синонимичных словоформ, которые могут быть использованы при определении их информативности (две синонимичные словоформы считаются двумя вхождениями лексемы в текст документа).

На первоначальном этапе информационная система может работать без сформированных словарей парадигм и синонимичных словоформ (т. е. с «пустыми» словарями).

2. Основные задачи системы

Конкретные задачи в системе решаются на основе реализации двух видов информационного поиска – поиска веб-страниц с выдачей их адресов, упорядоченных по убыванию информативности этих страниц, и фактографического поиска в полнотекстовых документах.

2.1. Индексирование документов и тематических корпусов текстов

Целью индексирования текста является приписывание ему совокупности ключевых слов с их весами (вес – это информативность слова). При индексировании используются абсолютные частоты слов в документе (если его объем достаточно большой) или в релевантном тексту тематическом корпусе текстов (если это краткое сообщение, т. е. объем текста небольшой), а также абсолютные частоты слов в полном корпусе текстов. Информативность слова вычисляется как отношение этих частот. При этом:

– частота слова в документе – это сумма частот всех словоформ, встречающихся в документе и являющихся словоизменениями исходной словоформы или ее синонимами, зафиксированными в словаре словоизменительных парадигм и в словаре синонимичных словоформ;

– частота слова в полном корпусе текстов – это сумма частот всех словоформ в полном корпусе текстов, в которой учтены словоизменения и синонимы.

Поисковый образ тематического корпуса текстов создается в виде набора ключевых слов с весами. Индексирование реализуется по аналогии с индексированием текстовых документов большого объема. В данном случае все тексты тематического корпуса объединяются и индексируются как единый текстовый документ.

2.2. Индексирование кратких сообщений и запросов пользователей

Краткое сообщение – это текстовый документ, объем которого не позволяет выявить статистические характеристики его словоформ. Для индексирования краткого сообщения используется релевантный ему тематический или динамический корпус текстов. Поисковым образом краткого сообщения считается поисковый образ найденного релевантного тематического корпуса текстов, из

которого исключены все словоформы, не содержащиеся в кратком сообщении (с учетом словоизменения и синонимии).

При информационном поиске в системе используются два основных типа запросов:

– свободно формулируемые запросы на естественном языке. Это традиционный тип запроса. При индексировании всем его словам приписывается информативность, равная 100%. Возможно также индексирование запроса с предварительным поиском наиболее релевантного ему тематического корпуса текстов. В этом случае ключевым словам запроса ставятся в соответствие весовые коэффициенты из словаря словоформ;

– запросы, формулируемые пользователем с применением специального графического интерфейса, где каждое слово запроса располагается на вертикальной шкале информативности. В зависимости от местоположения слова ему присваивается соответствующий весовой коэффициент.

Запрос индексируется аналогично индексированию краткого сообщения. Возможности подсистемы индексирования проиллюстрированы на рисунке 1.

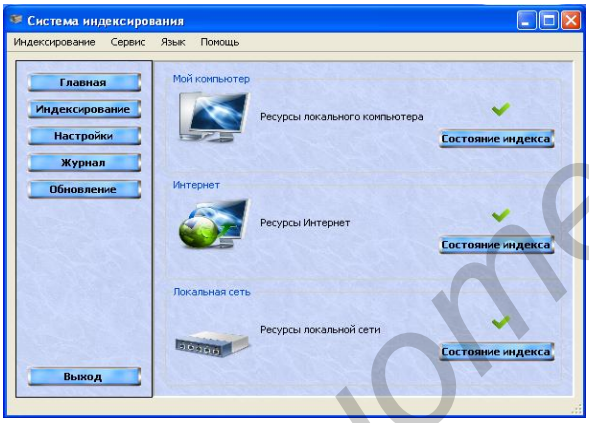

Рисунок 1 – Главное окно подсистемы индексирования

2.3. Поиск текстовых документов

Процесс поиска информации заключается в сравнении запросов пользователей с поисковыми образами проиндексированных документов. Поиску предшествует автоматическая коррекция запроса с целью адаптации системы к информационным потребностям пользователя. Коррекция реализуется следующим образом: на основе первоначального запроса создается динамический корпус текстов как подмножество полного корпуса; документы из динамического корпуса предъявляются пользователю, который исключает из него все непертинентные тексты; полученное в результате множество считается уточненным динамическим корпусом, на основе которого путем его индексирования формируется уточненное поисковое предписание. Процедура оценки пользователем пертинентности текстов может не проводиться. В этом случае для создания уточненного запроса используется исходный динамический корпус текстов. Главное окно подсистемы поиска представлено на рисунке 2.

2.4. Реферирование текстовых документов

Процесс реферирования включает следующие основные этапы: вычисление информативности слов и предложений реферируемого документа; разбиение текста на монотематические фрагменты и установление ситуативных связей между ними; вычисление информативности монотематических фрагментов; синтез реферата. Реферат строится из информативных предложений путем поиска релевантной информации в специальной системе словарей и последующего синтеза выходного текста. Алгоритм реферирования функционирует следующим образом.

Лингвистический процессор проводит синтаксический анализ реферируемого текста. В результате получаем упорядоченную совокупность синтаксических деревьев всех его предложений.

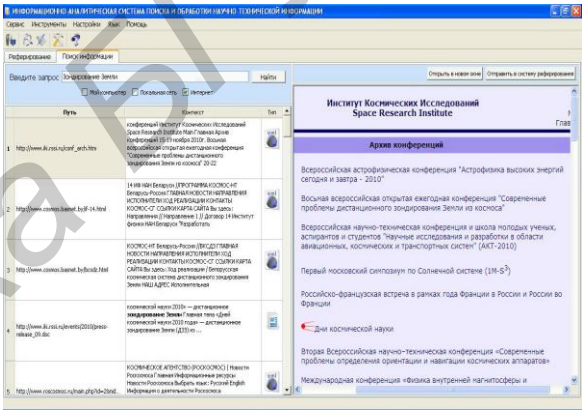

Рисунок 2 – Главное окно подсистемы поиска

Далее статистический анализатор определяет информативность каждого слова, т. е. эмпирическую вероятность того, что это слово извлечено из тематического корпуса текстов при условии, что оно уже извлечено из полного. Из полученной на этом шаге алгоритма совокупности синтаксических деревьев последовательно исключаются деревья (в порядке возрастания информативности слов) до получения требуемого объема будущего реферата. За информативность синтаксического дерева принимается максимальный из показателей информативности его слов. Далее из каждого оставшегося синтаксического дерева удаляются их неинформативные висячие поддеревья. Заключительными шагами алгоритма реферирования являются поиск в базе знаний адекватного синтаксического шаблона реферата, заполнение его слотов полученными на предыдущем шаге синтаксическими деревьями и синтез реферата на выходном языке. Главное окно подсистемы реферирования научно-технической информации изображено на рисунке 3.

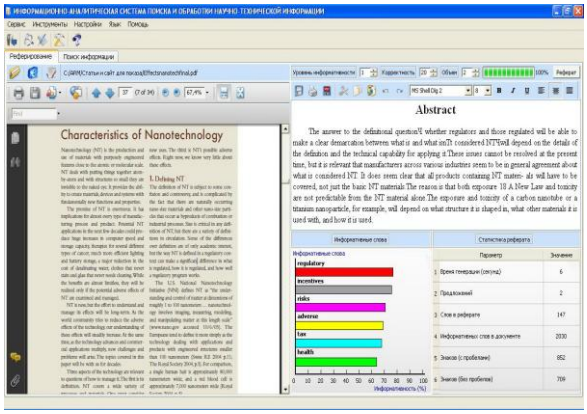

Рисунок 3 – Главное окно подсистемы реферирования

Заключение

Разработанная технология поиска и реферирования текстовой информации может быть использована в библиотеках, в информационно-аналитических отделах различных служб и организаций, которые осуществляют оперативный сбор и аналитическую обработку текстовых документов по различным предметным областям в Интернете, локальных сетях и на жестких или съемных дисках отдельных компьютеров. Созданный программный комплекс обеспечивает:

– индексирование, поиск и реферирование текстовых документов из различных информационных источников в многоязычной среде;

– многопоточность процессов индексирования, поиска и реферирования;

– поддержку наиболее распространенных форматов представления текстовых документов (html, shtml, doc, rtf, docx, pdf, txt) с возможностью подключения дополнительных форматов, таких как ppt, xls, wpd, hlp, odt и xml.

RETRIEVAL AND SUMMARIZATION OF TEXT INFORMATION IN A MULTILINGUAL ENVIRONMENT

Lipnitsky S.F., Mamchich A.A., Stepura L.V.

The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Republic of Belarus

lipn@newman.bas-net.by

lexamam@newman.bas-net.by

stepura@newman.bas-net.by

The technology of search and processing of scientific and technical information from different sources (Internet, LAN, user's hard disk) in a multilingual environment is considered. Information system architecture is proposed. Its main functions are presents: indexing, retrieval and automatic summarization of text documents.

Introduction

Automated system of retrieval and summarization of textual information in a multilingual environment was developed in UIIP NASB. Implemented in the system models and algorithms are used to enhance and effectively use information resources from various sources (Internet, LAN, hard drives of individual computer users).

Proposed in this paper approach to retrieving and summarization the text information, in contrast to existing ones, based on thematic and dynamic text corpora (set of texts on specific topics), which adapts the system to the task and the information needs of users, and independent software of the input language. Such a text corpora can be created under the projected pre-task and quickly formed after the request (dynamic text corpora).

Main Part

In order to improve the efficiency of indexing, retrieving and summarization documents from various information sources, in this paper we proposed an approach which, unlike existing methods, is based on using thematic corpora (collections of texts on specific topics) as a knowledge domain and specialized knowledge base dictionaries formed on their basis.

This technique provides the system adaptation to the task and the independence of a software complex from the input language. Text corpora can be created under the projected pre-task or formed directly on-line by combining sets of documents that are relevant to each particular text or a user query (we called them – dynamic corpora). This provides the adaptation to user information needs and gives a possibility to index and search not only full-text documents but short messages too, volumes of which are small and don't allow identifying their statistical characteristics.

The proposed algorithms are notable for universality, i. e. for the independence from topics. Information system adjustment to a particular data domain can be fully automated, it adds up to the creation of a respective thematic text corpus and an actualization of dictionaries of the knowledge database. At the same time, formation of the given dictionaries can completely be carried out in the hands off.

Conclusion

The algorithms presented in this paper can be used in various systems designed for processing and analyzing texts. The proposed technique of calculating the informativity of index terms can be used in automatic summarization systems for detecting informative wordforms in documents and for synthesizing connected summaries. With an appropriate selection of subjects and the hierarchical structure of a text corpus, it is possible to search regarding documents stylistic color (e.g. journalism, popular or scientific literature).