ПОИСК И РЕФЕРИРОВАНИЕ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ В ЛОКАЛЬНЫХ И ГЛОБАЛЬНЫХ КОМПЬЮТЕРНЫХ СЕТЯХ

А.Г. Буравкин
Ведущий научный сотрудник ОИПИ НАН Беларуси, кандидат технических наук, Республика Беларусь

С.А. Кореняко
Заведующий отделом ОИПИ НАН Беларуси, Республика Беларусь

С.Ф. Липницкий
Главный научный сотрудник ОИПИ НАН Беларуси, доктор технических наук, Республика Беларусь

Л.В. Степура
Научный сотрудник ОИПИ НАН Беларуси, Республика Беларусь

Объединенный институт проблем информатики Национальной академии наук Беларуси, lipn@newman.bas-net.by

Automated system of retrieval and summarization of textual information 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).

Введение

В докладе рассматривается автоматизированная система поиска и реферирования текстовой информации, 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].

1. Архитектура системы поиска и реферирования научно-технической информации

Функциональными компонентами системы поиска и реферирования текстовых документов являются три подсистемы:

– автоматизированное рабочее место (АРМ) эксперта-лингвиста;

– подсистема поиска текстовых документов;

– подсистема реферирования найденных документов.

АРМ эксперта-лингвиста – это инструментарий, предназначенный для автоматизированного создания и актуализации баз данных и знаний информационной системы. В базе данных эксперт-лингвист накапливает и классифицирует по тематическим разделам предметной области электронные варианты различных документов, на основе которых в базе знаний формируются тематические корпусы текстов для всех разделов предметной области и лингвистические словари базы знаний.

В состав подсистемы поиска научно-технической информации входят следующие информационно-программные средства: программы индексирования веб-страниц; программы поиска веб-страниц; программы поиска в полнотекстовых документах по содержанию; программы выявления информативных слов и предложений.

Основной функцией подсистемы реферирования является выявление в реферируемом тексте кортежа информативных предложений и синтез реферата. В состав подсистемы входят следующие основные структурные компоненты: программы поиска информативных предложений, которые выявляют в тексте информативные предложения и «высвечивают» их; программы синтеза рефератов из информативных предложений путем поиска релевантной информации в специальной системе словарей и последующего синтеза выходного текста.

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 [2].

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 пользователя индексируется аналогично индексированию краткого сообщения.

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 пользовате-

лем пертинентности текстов может не проводиться. В этом случае для создания уточненного запроса используется исходный динамический корпус текстов [3].

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. 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.

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;

– многопоточность процессов индексирования, поиска и реферирования;

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

Благодаря использованию тематических корпусов текстов достигается универсальность программного обеспечения системы, т. е. его независимость от используемых входных языков и предметной области.

Литература

1. Липницкий, С.Ф. Модель представления знаний в информационных системах на основе вербальных ассоциаций / С.Ф. Липницкий // Информатика. – 2011. – № 4. – С. 21–28.

2. Липницкий, С.Ф. Индексирование текстовой информации на основе моделирования вербальных ассоциаций / С.Ф. Липницкий // Информатика. – 2012. – № 3. – С. 94–102.

3. Липницкий, С.Ф. Моделирование информационного поиска на основе динамических корпусов текстов / С.Ф. Липницкий, А.А. Мамчич // Весці НАН Беларусі. Сер. фіз.-тэхн. навук. – 2011. – № 1. – С. 72–81.