УДК 004.912:002:63

ПРОЕКТ СИСТЕМЫ АВТОМАТИЗИРОВАННОГО РЕФЕРИРОВАНИЯ ЭЛЕКТРОННЫХ МАССИВОВ НАУЧНО-ТЕХНИЧЕСКИХ ПУБЛИКАЦИЙ ПО АГРАРНОЙ ТЕМАТИКЕ

Буравкин Алексей Геннадьевич,
Объединенный институт проблем информатики НАН Беларуси,
Минск, Беларусь
buralex@tut.by

Липницкий Станислав Феликсович,
Объединенный институт проблем информатики НАН Беларуси,
Минск, Беларусь
lipn@newman.bas-net.by

Степура Людмила Васильевна,
Объединенный институт проблем информатики НАН Беларуси,
Минск, Беларусь
stepura@newman.bas-net.by

Стрелкова Ирина Борисовна,
Государственное учреждение «Белорусская сельскохозяйственная библиотека им И.С. Лупиновича»
Национальной академии наук Беларуси,
Минск, Беларусь
irin-strelkova@yandex.ru

Представлены основные системотехнические параметры проекта информационной системы автоматизированного реферирования электронных массивов научно-технических публикаций по аграрной тематике. Реализация проекта позволит существенно сократить время аналитической обработки электронных архивов публикаций.

Ключевые слова: информационная система, автоматизированное реферирование, аналитическая обработка.

PROJECT OF AUTOMATIC SUMMARIZATION SYSTEM OF ELECTRONIC ARRAYS OF SCIENTIFIC AND TECHNICAL PUBLICATIONS ON AGRICULTURAL TOPICS

Alexey Burawkin,
The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
Minsk, Belarus
buralex@tut.by

Stanislaw Lipnicki,
The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
Minsk, Belarus
lipn@newman.bas-net.by

Lyudmila Stepura,
The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
Minsk, Belarus
stepura@newman.bas-net.by

Irina Strelkova,
State Institution «I.S. Lupinovich Belarus Agricultural Library»
of the National Academy of Sciences of Belarus,
Minsk, Belarus
irin-strelkova@yandex.ru

Presented basic parameters of the project of automatic summarization system of electronic arrays of scientific and technical publications on agricultural subjects. The project will significantly reduce the time analytical processing of electronic archives of publications.

Keywords: information system, automatic summarization, analytical processing.

Введение

Согласно результатам исследований компании IDC (International Data Corporation), объемы публикуемых данных еже-

годно удваиваются, причем доля полезной информации в этом стремительном потоке составляет лишь 35%. В связи с такой динамикой роста количества публикаций и с учетом значительного информационного шума специалисты вынуждены (с целью экономии времени) обращаться к вторичным документам – рефератам и аннотациям. Как реферирование, так и аннотирование текстовых документов используется для сокращения их объема при сохранении основного содержания. Однако результаты семантического сжатия первоисточников в реферате и аннотациях существенно отличаются.

В реферате кратко представлено содержание текстового документа, включающее основные фактические сведения и выводы без изложения субъективных взглядов на документ и его оценки. Реферат дополняет библиографическое описание публикации и дает первичное представление о ней.

Эффективными средствами организации работы специалистов с текстовыми документами большого объема являются системы их автоматизированного реферирования. Процесс реферирования в таких системах включает, как правило, три этапа. На первом этапе анализируется реферируемый текст, на втором выявляются информативные слова, предложения и фрагменты текста, а на третьем синтезируется его реферат. Таким образом, автоматизированное реферирование относится к классу задач аналитико-синтетической обработки текстовой информации, решение которых связано с использованием специализированных баз знаний.

Существующие подходы к автоматическому реферированию основаны главным образом на различных эвристических, статистических и лингвистических методах, что не позволяет достичь необходимой эффективности реферирования (например, адаптивности системы к предметной области пользователей, независимости программного обеспечения от входных языков, сравнительно простых и удобных средств создания баз знаний и словарей).

Одним из направлений повышения эффективности систем автоматического реферирования является их интеллектуализация, т.е. придание им способности «компьютерного понимания» реферируемых текстов и адаптации на этой основе алгоритмов реферирования к обрабатываемым данным и изменяющимся условиям взаимодействия пользователя с системой. Для решения задач интеллектуализации необходим подход, обеспечивающий возможность модели-

рования и алгоритмизации всех информационных процессов аналитико-синтетической переработки информации в системе реферирования в рамках единой теории. В настоящее время такие подходы отсутствуют в связи с преобладанием эвристических методов исследования проблемы.

Решение задач интеллектуализации систем автоматического реферирования обеспечит возможность региональной интеграции информационных служб различных организаций и учреждений образования, в том числе и сельскохозяйственного профиля, за счет использования новых проектных решений при разработке программного и лингвистического обеспечения.

Информационные системы с функцией реферирования различают главным образом по виду обрабатываемых данных (полнотекстовые документы или фактографические сведения). Методы обработки фактографических данных в настоящее время в значительной степени разработаны. Наиболее известны из них технологии оперативной аналитической обработки (OLAP – On-Line Analytical Processing) и интеллектуального анализа данных (Data Mining). Существующие методы анализа полнотекстовых документов не отличаются большими функциональными возможностями и сводятся в основном к тематическому рубрицированию текстов и подсчету статистики встречаемости слов и словосочетаний.

Существуют встроенные программные модули (например, функция AutoSummarize офисного пакета Microsoft Office), предназначенные для построения реферата только на одном языке (например, на английском) и использующие простые статистические и позиционные алгоритмы.

Интерес представляют программные разработки Copernic Summarizer, Inxight Summarizer, Oracle Text, которые являются многоязычными системами реферирования, предназначенными для обработки текстов как на европейских языках (Copernic Summarizer), так на восточных – японском, китайской и корейском языках (Inxight Summarizer, коммерческий программный продукт для реферирования текстов, в основе работы которого заложены лингвистические алгоритмы, разработанные Исследовательским центром Ксерокс в Пало Альто). Проведение тематического анализа текстов на английском языке может быть выполнено с помощью средств Oracle Text (разработка компании ФОРС, Москва, Россия). В ходе обработки текст каждого документа подвергается процедурам лингви-

стического и статистического анализа, в результате чего определяются его ключевые темы и строится общее резюме – реферат.

Внимания заслуживает программа реферирования «МЛ Аннотатор» (разработка российской компании МедиаЛингва), предоставляющая разработчикам программного обеспечения комплект технической документации и набор специализированных инструментов (SDK – Software Development Kit) для автоматизированного аннотирования документов любого объема и степени сложности на русском и английском языках. Данная программа содержит гибкие средства для настройки режимов и параметров реферирования – от выделения ключевых терминов до объема аннотаций. Эта разработка позволяет вычислять критерии значимости и семантической независимости для предложений входного текста с использованием различных интеллектуальных алгоритмов на основе специальных вероятностных моделей и машинной морфологии русского языка в виде набора словарей.

Цель и назначение проекта

Целью проекта является создание алгоритмов, программных и информационно-лингвистических средств системы автоматизированного реферирования многоязычных электронных массивов научно-технических публикаций в многоязычной среде для ввода в эксплуатацию в Государственном учреждении «Белорусская сельскохозяйственная библиотека им И.С. Лупиновича» Национальной академии наук Беларуси (БелСХБ).

Система автоматизированного реферирования многоязычных электронных массивов текстовой информации предназначена для накопления и семантического сжатия публикаций сельскохозяйственной научной тематики, представленных в электронном формате. В настоящее время в БелСХБ накоплен архив публикаций прошлых лет, не потерявших актуальности и подлежащих реферированию. Применение системы автоматизированного реферирования позволит существенно сократить время аналитической обработки архивов публикаций.

Основными направлениями использования создаваемой системы являются:

1) Ретроспективное – повышение эффективности аналитической обработки материалов, полученных в результате ретроспек-

тивной конверсии (подготовка обзоров различных видов научных изданий сельскохозяйственной тематики).

2) Перспективное – повышение эффективности аналитической обработки стремительно растущих объемов интернет-публикаций [1].

Общая характеристика результатов проекта

В результате разработки системы автоматизированного реферирования многоязычных электронных массивов научно-технических публикаций предполагается создание следующей научно-технической продукции:

– программ создания и ведения тематических корпусов текстов на белорусском, русском и английском языках;

– программ создания и ведения лингвистических словарей;

– программ вычисления информативности слов в текстовых документах;

– программ автоматического реферирования текстовых документов;

– программ реализации веб-интерфейса;

– тематических корпусов текстов на белорусском, русском и английском языках;

– лингвистических словарей;

– эксплуатационной документации.

Программно-информационный комплекс, который будет создан в результате реализации проекта, будет обладать следующими особенностями:

– комплекс будет многопользовательским;

– он будет иметь модульную структуру и допускать возможность развития и модернизации его частей;

– система реферирования будет функционировать в многоязычной среде;

– будет реализована функция распределения доступа пользователей к ресурсам системы;

– система реферирования позволит создавать и актуализировать корпусы текстов по различным научным тематическим направлениям в многоязычной среде;

– система обеспечит автоматическое вычисление информативности слов и предложений в текстовых документах в многоязычной среде;

– система обеспечит предварительное задание пользователем объема реферата (количества предложений в нем).

К основным отличительным особенностям данного проекта относятся:

– использование при разработке программ вычисления информативности слов и предложений реферируемых текстов новых теоретических результатов, полученных исполнителями проекта [2]. Эти результаты обеспечивают устойчивость алгоритмов к орфографическим ошибкам, что особенно актуально при реферировании текстов, полученных в результате ретроспективной конверсии;

– архитектура системы, включающая служебную, административную и сервисную части;

– способы реализации, использующие современную инструментальную среду для веб-разработок.

Заключение

Белорусская сельскохозяйственная библиотека заинтересована в реализации данного проекта и готова выступить в качестве экспериментальной площадки. Программный комплекс автоматизированного реферирования мультиязычных электронных массивов научно-технических публикаций в перспективе позволит автоматизировать процесс аналитической обработки, в т.ч. реферирования, электронных документов библиотек, издательств и др. организаций. Сегодня поток электронных национальных документов незначителен, поэтому есть время для производства и апробации таких интеллектуальных инструментов, как Программный комплекс автоматизированного реферирования мультиязычных электронных массивов научно-технических публикаций. Это уже европейский уровень аналитической научной обработки публикаций и подготовки рефератов.

Список использованных источников:

1. Липницкий, С. Ф. Информационная система интернет-мониторинга публикаций: функции и структура / С. Ф. Липницкий, Л. В. Степура, А. Г. Буравкин // Библиотеки в информационном обществе: сохранение традиций и развитие новых технологий : доклады Междунар. науч. конф., Минск, 3–4 дек. 2014 г. / ГУ «Белорусская сельскохозяйственная библиотека им. И. С. Лупиновича» НАН Беларуси ; редкол.: В. В. Юрченко [и др.] ; рец.: Р. Б. Григянец, С. В. Зыгмантович. – Минск : Ковчег, 2014. – С. 47–53.

2. Липницкий, С. Ф. Поиск и реферирование текстовой информации в многоязычной среде / С. Ф. Липницкий, А. А. Мамчич, Л. В. Степура // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS–2013) : материалы III Междунар. научн.-техн. конф., Минск, 21–23 февр. 2013 г. – Минск : БГУИР, 2013. – С. 229–232.