Поиск и обработка информации на базе мультилингвистических тезаурусов

Зеленков П.В., Ковалев И.В., Карасева М.В., Рогов С.С.

В настоящее время разработано множество моделей и алгоритмов для представления информации в информационных системах. Частным случаем подобных систем являются информационно-управляющие системы, корпоративные информационные системы и системы поддержки принятия решения. Однако большинство моделей распределенных систем строятся на основе одноязычного представления информации или учитывают многоязычность неявно.

Одним из перспективных направлений является применение предметных словарей или тезаурусов. Необходимо отметить, что в современных системах подобные словари-тезаурусы очень редко встречаются представленными в мультилингвистической частотной реализации. Авторами статьи в рамках предлагаемых моделей применяются тезаурусы, выполненные на основе мультилингвистической технологии для проведения поисковой процедуры в информационных системах.

Авторский подход направлен на решение проблемы многоязычного представления информации в информационно-управляющих системах. В современных условиях даже небольшие корпоративные информационные системы работают в мультилингвистическом режиме.

В работе программный модуль выполняет основную работу по поиску, ранжированию и определению уровня релевантности документов путем использования метапоисковых мультилингвистических алгоритмов обработки информации. Для этого необходимо определить параметры процесса поиска. К ним относятся функции выбора предметной области и настройки языковых множеств, в рамках которых необходимо производить поиск.

Кроме того, необходимо отдельно показать возможность работы со строкой поиска информации, как в Internet, так и корпоративной сети. Согласно предлагаемому авторами подходу работа с поисковой строкой может проводиться в двух режимах:

- режим ручного ввода строки поиска;
- режим автоматизированного формирования строки поиска.

При ручном режиме система проверяет наличие введенных термов в частотном мультилингвистическом тезаурусе и в случае отсутствия терма в словаре пользователю предлагается ввести поисковую строку с изменением термов в строке поиска.

Рассмотрим процесс формирования запроса по заданной предметной области при автоматическом режиме. Так как модуль поиска информации основан на применении частотных мультилигнвистических тезаурусов, то, исходя из частотных характеристик терминов, можно сформировать поисковую строку. Пользователь может корректировать ее или дополнять.

В современных корпоративных системах храниться, как правило, мультилингвистическая информация, однако пользователь поискового модуля не может знать всех языков, представленных в сети. Поэтому необходимо учитывать конкретность указания языковых множеств, необходимых пользователю.

При завершении процесса формирования поисковой строки и указания языков поиска, необходимо приступать непосредственно к поисковой процедуре. В результате происходит последовательный опрос всех информационных корпоративных ресурсов и формируется массив ссылок на интересующие пользователя документы, а также происходит разбиение всего множества ссылок по принципу принадлежности к языковому множеству.

Также, пользователь может увидеть дополнительную информацию, которая учитывается при ранжировании документов и определении уровня релевантности каждого найденного документа:

- заголовок документа;

– объем документа;
– количество найденных термов в документе.

На втором шаге происходит определение уровня релевантности и ранжирование мультилингвистического массива ссылок. Здесь пользователю предоставляется дополнительная информация уже другого рода:

– уровень релевантности найденного документа;
– общий вес релевантных термов в документе.

Третий шаг - это непосредственный просмотр найденных документов. Необходимо отметить, что на данном этапе можно не только просмотреть документ, но и получить о нем расширенную информацию.

Таким образом, предлагаемый авторами модуль поиска и обработки информации в корпоративных системах поддержки принятия решения полностью удовлетворяет современным требованиям систем подобного уровня и позволяет решить проблему организации, хранения и обработки информации в современных распределенных мультилингвистических корпоративных системах поддержки принятия решений.

Кроме того, представленные мультилингвистические модели позволяют составить мультилингвистические ответы даже на одноязычные запросы более гибко с учетом неопределенности описания как мультилингвистических документов, так и запросов.