МЕТОДИКА СОЗДАНИЯ ТЕГ-ОРИЕНТИРОВАННОЙ БАЗЫ ДАННЫХ

Захаревская А.В., Липчанский М.В.
Национальный технический университет «Харьковский политехнический институт», г. Харьков

Основные идеи современной информационной технологии базируются на концепции баз данных. Согласно этой концепции, основой информационной технологии являются данные, которые должны быть организованы в базах данных с целью адекватного отображения изменяющегося реального мира и удовлетворения информационных потребностей пользователей. Для эффективной работы с информацией такого огромного объема нужна высокая степень ее упорядочения. Современные системы управления базами данных предоставляют развитые средства для организованного доступа к информации. Но на сегодняшний момент появляется все больше задач, которые требуют упорядочивания большого объема текстовых данных, написанных природным языком, на основе слабо структурированных нечетких правил.

В качестве примера неструктурированной формы можно привести: связный текст (т.е. документ на естественном языке – на литературном, официально-деловом и т.д.), графические данные в виде фотографий, картинок и прочих неструктурированных изображений.

Такие данные поступают хаотично и из разных источников, имеют много связей, могут относиться к разным элементам структуры, их фрагменты относятся к разным сущностям объекта – все это причины, из-за которых создать одну хорошую структуру для этих данных достаточно тяжело.

Во время решения таких задач возникает такое техническое противоречие: в момент создания базы данных все данные должны быть строго формализованы с использованием традиционных баз данных. С другой стороны, известно, что этот класс задач меняется и уточняется по ходу развития проекта. То есть традиционные базы данных удобны для сохранения в общем, но часто их структура не известна заранее и уточняется в процессе сбора этих данных. Возникает сложность в формировании структуры базы данных, которая не будет требовать значительной траты времени на пояснения всех отношений, которые необходимы для корректного построения реляционной модели. Чтобы вывести данные в таком виде необходимо извлечь данные из нескольких файлов, скомбинировать их и представить их вместе. Причина возникшего затруднения в том, что в системах обработки файлов связи между записями не представлены в явной форме и не обрабатываются.

Решение выше обозначенного противоречия является актуальной на данный момент задачей, которая требует научного подхода и инновационных решений, и есть темой данного доклада. Поэтому цель данной работы – разработка теоретических основ и инструментальных способов создания тег-ориентированной базы данных для слабо структурированных заданий. Это позволит сохранять данные сразу с необходимым идентификатором или их группой, быстро находить необходимую информацию, а после – отобразить ее в необходимом виде.