# Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества

# Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation

***Материалы конференции***
***Conference Program***

Том 1
Volume 1

г. Евпатория, Республика Крым, Украина
10-18 июня 1995 г.
Eupatory, Republic of Crimea, Ukraine
10-18 June 1995

# Проблемы эксплуатации больших баз данных: сжатие данных

# Operations with Large Databases: Data Compression

**Мазов Н.А.**
*Объединенный институт геологии, геофизики и минералогии Сибирского отделения РАН, Новосибирск, Россия*

**Mazov N.A.**
*Joint Institute of Geology, Geophysics and Mineralogy, Siberian branch of Russian Academy of Sciences, Novosibirsk, Russia*

*The paper is devoted to data compression and to algorithms used for this task. Redundancy of data, that are processed in information systems, may be detected and revealed by studying their statistical and semantic features. The author examines various data compression techniques for STI databases, that are run in the CDS/ISIS/M software package environment.*

В настоящее время работа большинства информационных систем связана с хранением и использованием очень больших объемов информации. Большинство баз данных (БД), используемых в таких системах, содержит сотни мегабайт информации. Массивы таких размеров увеличивают стоимость эксплуатации таких систем; кроме того, при эксплуатации БД возникают трудности из-за огра-

ниченного размера внешней памяти имеющейся в распоряжении служб НТИ и библиотек. Поэтому очевидно, что вопросам сжатия данных и алгоритмам реализующих их, в последнее время проявляется достаточный интерес. Основная цель сжатия данных заключается в том, чтобы максимально уменьшить исходную длину записи в БД. На практике оно применяется в основном для:

— хранения данных в архивных файлах;

— передачи данных по сетям ЭВМ;

— хранения массивов массивов БД во внешней памяти с прямым доступом.

Если первые два аспекта неплохо изучены и имеется достаточное количество программных средств для их реализации, то последний, на наш взгляд, изучен недостаточно, особенно для систем НТИ.

Различные эксперименты показывают, что представление информации в БД НТИ весьма избыточно, ввиду их специфичности и ограниченности лексики. Использование этого свойства позволяет сократить исходную информацию как мимнимум наполовину и тем самым существенно повысить эффективность информационных систем, а при ограниченных ресурсах внешней памяти значительно повысить их производительность.

На практике обычно при оценке различных методов сжатия данных учитывается в основном три фактора, которые достаточно полно позволяют их оценивать:

— коэффициент сжатия, который определяется отношением разности длин исходного и выходного сообщений к длине исходного сообщения;

— время необходимое для кодирования/декодирования сообщения;

— устойчивость коэффициента сжатия, т.е. его зависимость от изменения характера данных, используемых в качестве исходных сообщений.

При эксплуатации БД НТИ вторичной информации как правило приходится иметь дело с записями, длина которых не превышает десятка килобайт, а сами данные однородны по наполнению и структуре. Помимио этого в таких данных наблюдается высокая избыточность исходной информации. Поэтому последние два фактора не так актуальны при исследовании вопроса о сжатии данных. Это показывает и опыт эксплуатации БД, для которых применяется сжатие данных. Это такие БД Института научной информации США, как Current Contents on Disk и Science Citation Index и др.

На наш взгляд наиболее актуальным является первый фактор — коэффициент сжатия.

Сжатию данных должен предшествовать подробный анализ характеристик БД. При помощи анализа статистических и семантических свойств обрабатываемых в информационных системах данных становится возможным выявление и сокращение избыточности, поскольку эффективное сжатие невозможно без предварительного анализа данных и как следствие выбора того или иного метода сжатия.

Основным недостатком многих статей посвященных различным методам сжатия является недостаточно подробное описание экспериментов и характера обрабатываемых данных.

Автором были сделаны попытки исследования различных методов сжатия для БД НТИ, эксплуатируемых в Институте под управлением CDS/ISIS/M.

В ходе исследований были реализованы соответствующие программы, которые позволили:

— достаточно полно произвести статистический и семантический анализ используемых полей БД НТИ;

— получить частотные распределений отдельных символов, би-грамм и триграмм для ряда полей БД НТИ;

— получить оценки об эффективности применения различных методов сжатия по коэффициенту сжатия;

— оценить поведение CDS/ISIS/M при использовании сжатых данных.

В докладе подробно обсуждается применение различных подходов к сжатию данных в БД НТИ, эксплуатируемых под управлением CDS/ISIS/M.