Chapter 2
Practical Approaches to the Many–Answer Problem

Mounir Bechchi
LINA-University of Nantes, France

Guillaume Raschia
LINA-University of Nantes, France

Noureddine Mouaddib
LINA-University of Nantes, Morocco

ABSTRACT
Database systems are increasingly used for interactive and exploratory data retrieval. In such retrievals, users’ queries often result in too many answers, so users waste significant time and efforts sifting and sorting through these answers to find the relevant ones. This chapter first reviews and discusses several research efforts that have attempted to provide users with effective and efficient ways to access databases. Then, it focuses on a simple but useful strategy for retrieving relevant answers accurately and quickly without being distracted by irrelevant ones. Generally speaking, the chapter presents a very recent but promising approach to quickly provide users with structured and approximate representations of their query results, a must have for decision support systems. The underlying algorithm operates on pre-computed knowledge-based summaries of the queried data, instead of raw data themselves. Thus, this first-citizen data structure is also presented in this chapter.

1. INTRODUCTION
With the rapid development of the World Wide Web, more and more accessible databases are available online; A July 2000 study (Bergman, 2001) estimated 96000 relational databases were online and the number increased by seven times in 2004 (Chang, He, Li, Patel & Zhang, 2004). The increased visibility of these structured data repositories made them accessible to a large number of lay users, typically lacking a clear view of their content, moreover, not even having a particular item in mind. Rather,
they are attempting to discover potentially useful items. In such a situation, user queries are often very broad, resulting in too many answers. Not all the retrieved items are relevant to the user. Unfortunately, she/he often needs to examine all or most of them to find the interesting ones. This too-many-answers phenomenon is commonly referred to as information overload - “a state in which the amount of information that merits attention exceeds an individual’s ability to process it” (Schultz & Vandenbosch, 1998).

Information overload often happens when the user is not certain of what she/he is looking for, i.e., she/he has a vague and poorly defined information need or retrieval goal. Thus, she/he generally poses a broad query in the beginning to avoid exclusion of potentially interesting results and next, she/he starts browsing the answer looking for something interesting. Information overload makes it hard for the user to separate the interesting items from the uninteresting ones, thereby leading to potential decision paralysis and wastage of time and effort. The dangers of information overload are not to be underestimated and are well illustrated by buzzwords such as *Infoglut* (Allen, 1992), *Information Fatigue Syndrome* (Lewis, 1996), *TechnoStress* (Weil & Rosen, 1997), *Data Smog* (Shenk, 1997), *Data Asphyxiation* (Winkle, 1998) and *Information Pollution* (Nielsen, 2003).

In the context of relational databases, *automated ranking* and *clustering of query results* are used to reduce information overload. Automated ranking-based techniques first seek to clarify or approximate the user’s retrieval goal. Then, they assign a score to each answer, representing the extent to which it is relevant to the approximated retrieval goal. Finally, the user is provided with a ranked list, in descending order of relevance, of either all query results or only a top-k subset. In contrast, clustering-based techniques assist the user to clarify or refine the retrieval goal instead of trying to learn it. They consist in dividing the query result set into dissimilar groups (or clusters) of similar items, allowing users to select and explore groups that are of interest to them while ignoring the rest. However, both of these techniques present two major problems:

- the first is related to *relevance*. With regard to automated ranking-based techniques, the relevance of the results highly depends on their ability to accurately capture the user’s retrieval goal, which is not an obvious task. Furthermore, such techniques also bring the disadvantage of match homogeneity, i.e., the user is often required to go through a large number of similar results before finding the next different result. With regard to clustering-based techniques, there is no guarantee that the resulting clusters will match the meaningful groups that a user may expect. In fact, most clustering techniques seek to only maximize some statistical properties of the clusters (such as the size and compactness of each cluster and the separation of clusters relative to each other);
- the second is related to *scalability*. Both ranking and clustering are performed on query results and consequently occur at query time. Thus, the overhead time cost is an open critical issue for such a posteriori tasks.

To go one step beyond an overview of these well-established techniques, we investigate a simple but useful strategy to alleviate the two above problems. Specifically, we present an efficient and effective algorithm coined Explore-Select-Rearrange Algorithm (*ESRA*) that provides users with hierarchical clustering schemas of their query results. *ESRA* operates on pre-computed knowledge-based summaries of the data, instead of raw data themselves. The underlying summarization technique used in this work is the SAINTETIQ model (Raschia & Mouaddib, 2002; Saint-Paul, Raschia, & Mouaddib, 2005), which is a domain knowledge-based approach that enables summarization and classification of structured data stored into a database. Each node (or summary) of the hierarchy provided by *ESRA* describes a subset of