A Method for Similarity-Based Grouping of Biological Data

Vaida Jakonienė, David Rundqvist, and Patrick Lambrix

Department of Computer and Information Science
Linköpings universitet, SE-581 83 Linköping, Sweden

Abstract. Similarity-based grouping of data entries in one or more data sources is a task underlying many different data management tasks, such as, structuring search results, removal of redundancy in databases and data integration. Similarity-based grouping of data entries is not a trivial task in the context of life science data sources as the stored data is complex, highly correlated and represented at different levels of granularity. The contribution of this paper is two-fold. 1) We propose a method for similarity-based grouping and 2) we show results from test cases. As the main steps the method contains specification of grouping rules, pairwise grouping between entries, actual grouping of similar entries, and evaluation and analysis of the results. Often, different strategies can be used in the different steps. The method enables exploration of the influence of the choices and supports evaluation of the results with respect to given classifications. The grouping method is illustrated by test cases based on different strategies and classifications. The results show the complexity of the similarity-based grouping tasks and give deeper insights in the selected grouping tasks, the analyzed data source, and the influence of different strategies on the results.

1 Introduction

During the last decade an enormous amount of biological data has been generated and techniques and tools to analyze this data have been developed. Many of these tools use data clustering and classification techniques. For instance, these techniques are used to find similar sequences for predicting the functionality of new sequences [GH04], to find correlated genes based on microarray data [SS02], or to classify publications according to an ontology to locate relevant documents faster [DS05]. A basic task underlying these approaches is the computation of a similarity value between objects. Different techniques are developed to compute a similarity value between objects based on the object types. For instance, edit distance [Lev66] and n-gram [PPF95] are well-established techniques to define similarity between strings, while BLAST [AGMML90] can be used to define a similarity measure between DNA or protein sequences. Recently, a number of projects discussed methods to compute semantic similarity over terms in a Gene Ontology (GO) ontology (e.g. [CSC05] and [SFSZ05]). The similarity between GO terms can be used to compute a similarity between data entries that are annotated with these GO terms [LSBG03].

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 136–151, 2006.

Data entries in biological data sources are often complex and store different types of information. Although most of the research has focused on organizing the data based on aspects, such as sequence similarity and function, we need to analyze data using different aspects and from different points of view to obtain deeper insights in the characteristics of the data and to discover new knowledge. This means that we need to be able to organize the data based on different attributes or different combinations of attributes. [KLKTB04] illustrates how a combination of attributes could be used to find data entries describing the same protein. In this case, search on sequence similarity is complemented with the analysis of sequence length, organism and the data source where the sequence was originally submitted. In this paper we use the term *grouping* to refer to the task of organizing the data according to a certain aspect or a combination of aspects. Further, we concentrate on the task of *similarity-based grouping*. During similarity-based grouping the analyzed data entries are compared with respect to a selected subset of attributes, and similarity functions that are relevant to the attributes are used to compute the similarity of the stored values.

Grouping of data entries in one or more data sources is an operation underlying many different data management tasks. Grouping can be used to structure and visualize search results in a convenient way for the user. This is especially important when large data sources are studied. The possibility to get an overview over the data may lead to the discovery of new knowledge or may allow biologists to locate the information of interest faster. The identification of similar data entries and their grouping are core operations when performing data cleaning activities [HGPWW04]. The identified groups of similar data entries can be further analyzed and merged into a single data entry. In the context of data integration, techniques underlying grouping are important to correlate data entries at different data sources. The grouping task can be narrowed to the duplicate detection task, where it is required that matched data entries represent the same real-world object. Duplicate detection can be both used for data cleaning [KLKTB04] and for data integration [BBBDN05].

A number of aspects influence the quality of the grouping results: the quality of the data sources, the selection of the grouping attributes and the algorithms implementing the grouping procedure. In some cases, given a grouping task, it can be difficult to decide on which attributes to perform grouping. Also, different sets of attributes may seem relevant to the grouping task, but lead to varying quality of the results [KLKTB04]. Further, suitable algorithms need to be selected to compute the similarity between data entries and to organize similar data entries into groups. Many methods exist, but it is often not clear which methods perform best for which grouping tasks. The study of the properties, and the evaluation and the comparison of the different aspects that influence the quality of the grouping results, would give us valuable insight into the best way to use the grouping procedures. It would also lead to recommendations on how to improve the current procedures and develop new procedures. To be able to perform such studies and evaluations we need environments that allow us to compare and evaluate different grouping procedures.