Journal of Visual Languages and Computing 000 (2017) 1–12

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Rainbow boxes: A new technique for overlapping set visualization and two applications in the biomedical domain[☆]

Jean-Baptiste Lamy[a,b,*], Hélène Berthelot[a,b], Coralie Capron[a,b], Madeleine Favre[c]

[a] LIMICS, Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny, France
[b] INSERM UMRS 1142, UPMC Université Paris 6, Sorbonne Universités, Paris, France
[c] Dept. of primary care, Université Paris Descartes, Société de Formation Thérapeutique du Généraliste (SFTG), Paris, France

ARTICLE INFO

Article history:
Received 10 August 2017
Accepted 27 September 2017
Available online xxx

Keywords:
Overlapping set visualization
Information visualization
Knowledge visualization
User study
Drug properties
Gene annotation

ABSTRACT

Overlapping set visualization is a well-known problem in information visualization. This problem considers elements and sets containing all or part of the elements, a given element possibly belonging to more than one set. A typical example is the properties of the 20 amino-acids. A more complex application is the visual comparison of the contraindications or the adverse effects of several similar drugs. The knowledge involved is voluminous, each drug has many contraindications and adverse effects, some of them are shared with other drugs. Another real-life application is the visualization of gene annotation, each gene product being annotated with several annotation terms indicating the associated biological processes, molecular functions and cellular components.

In this paper, we present rainbow boxes, a novel technique for visualizing overlapping sets, and its application to the presentation of the properties of amino-acids, the comparison of drug properties, and the visualization of gene annotation. This technique requires solving a combinatorial optimization problem; we propose a specific heuristic and we evaluate and compare it to general optimization algorithms. We also describe a user study comparing rainbow boxes to tables and showing that the former allowed physicians to find information significantly faster. Finally, we discuss the limits and the perspectives of rainbow boxes.

1. Introduction

Overlapping set visualization is a well-known field in information visualization [1]. Several elements are considered, as well as sets containing all or part of these elements. The sets are *overlapping, i.e.* a given element can belong to more than one set. The objective of the visualization is to show clearly which elements belong to a given set, which sets include a given element but also to help answer more complex questions, for instance involving the intersection or disjointness of several sets, and to elicit new insights, such as finding similarities between elements or sets.

A typical and simple example of overlapping sets is the "*amino-acid properties*" problem. There are 20 amino-acids (*e.g.* Alanine, Proline) often abbreviated by their 3-letter codes (Ala, Pro) or by their 1-letter code (A, P). Several amino-acids share some physical or chemical properties, such as their *small* size, the presence of an *aromatic* cycle or a *positive* electric charge. A given amino-acid can have zero, one or several properties (*e.g.* Histidine has both an *aromatic* cycle and a *positive* charge). Thus the amino-acids can be considered as *elements* and the properties as *overlapping sets* of these elements. About 10 such properties are usually considered. In addition, properties are not independent from each other: for example, it is obvious that all *tiny* amino-acids are also *small*, and amino-acids with an *aromatic* cycle cannot be *small* (because the aromatic cycle is a big chemical structure). A good visualization is expected to show clearly which are the properties of a given amino-acid and which amino-acids share a given property, but also to facilitate the discovery of new knowledge about the amino-acids (*e.g.* Tyrosine and Tryptophan share the same properties and thus they possibly exhibit similar biological behaviors) and their properties (*e.g.* the relation between the *small* and *aromatic* properties detailed above).

More complex, real-life, overlapping set visualization problems exist. A first example is *the comparison of drug properties*. Drugs have many properties such as indications, contraindications, interactions, adverse effects, *etc.* A contraindication is a situation in which a given drug should be avoided (relative contraindication) or cannot be prescribed (absolute contraindication). The situation

[☆] This work was supported by the French Drug Agency (ANSM, Agence Nationale de Sécurité du Médicament et des produits de santé) through the VIIIP project [grant number AAP-2012-013].

[*] Corresponding author.

E-mail addresses: jibalamy@free.fr (J.-B. Lamy), helene.berthelot@orange.fr (H. Berthelot), coralie.capron@hotmail.fr (C. Capron), mfavre89@gmail.com (M. Favre).

https://doi.org/10.1016/j.jvlc.2017.09.003
1045-926X/© 2017 Elsevier Ltd. All rights reserved.

usually corresponds to a disease (*e.g.* “this drug is contraindicated with diabetes”) or a patient characteristic (*e.g.* “this drug is contraindicated for children”). An adverse effect is an undesired effect caused by a given drug (*e.g.* vomiting). Adverse effects are characterized by the seriousness of the effect and its frequency.

Drug properties are listed in official textual documents called summary of product characteristics (SPCs) and then gathered into drug databases. The user interfaces of these databases allow a physician to consult the properties of a single drug, but not to compare several drugs with the same indication. The VIIIP (Integrated Visualization of Information about Pharmaceutical Innovation, founded by the French drug agency) research project aims at facilitating the comparison of new drugs with the older similar drugs, and proposing visual interface for this task. However, the clear and concise presentation of the properties of a single drug is already difficult, and thus the visual comparison of the properties of 2–10 drugs is a real challenge. The comparison of drug properties can be expressed as an overlapping set visualization: drugs can be considered as *elements*, and their properties (*e.g.* contraindications or adverse effects) as *sets* including all drugs sharing the property[1]. For example, we might consider the set of drugs contraindicated with hypertension, or the set of drugs causing nausea (a common adverse effect).

A second complex example is *the visualization of gene annotation*. Gene annotation consists of associating one or more annotation terms to each gene products (usually proteins). One of the major resource for annotating genes is Gene Ontology (GO) [2]. GO includes three categories of annotation terms: cellular components (*e.g.* cytoplasm, nucleus), molecular functions (*e.g.* catalysis) and biological processes (*e.g.* lipid metabolism, cellular death). These terms are organized using hierarchical relations (is-a, part-of) and non-hierarchical relation (regulates). The GO database also includes the list of known gene products for many species, and the terms associated with each gene product. The visualization of the annotation of a set of gene products, for example a given family of protein, is difficult. This problem can be expressed as an overlapping set visualization: genes can be considered as *elements*, and annotation terms as *sets* including all genes associated with the term. For example, we might consider the set of gene products located in the cytoplasm or the set of genes participating in the lipid metabolism.

The objective of this paper is (1) to present *rainbow boxes*, a novel technique for visualizing overlapping sets that we developed initially for comparing drug properties, (2) to describe the user study we performed for evaluating this technique, and (3) to propose and evaluate a heuristic for solving the combinatorial optimization problem behind rainbow boxes. Rainbow boxes aim at presenting relatively small datasets, typically involving 2–25 elements and 5–100 sets, and at helping the discovery of classes of similar elements or sets.

The rest of the paper is organized as follows. Section 2 presents the state of the art in overlapping set visualization. Section 3 describes rainbow boxes on the simple amino-acid example. Section 4 describes the use of rainbow boxes for comparing drug properties. Section 5 describes the application of rainbow boxes to the visualization of gene annotation. Section 6 describes a user study that has been conducted to evaluate rainbow boxes in the drug comparison application. Section 7 evaluates the heuristic we propose for optimizing rainbow boxes and compares the heuristics with other algorithms. Finally, Section 8 discusses the results and proposes perspectives.

2. Related works

Alsakallah et al. recently reviewed the various methods proposed for overlapping set visualization [1]. They distinguished 6 approaches: (1) Euler / Venn diagrams and their variants, (2) overlays on a map, (3) node-link diagrams, (4) matrix-based techniques, (5) aggregation-based techniques, and (6) scatterplot-based techniques.

Here, we propose a classification in four main different approaches, grouping together Alsakallah's category 1, 2 and 6 since they all use the position of elements for indicating the sets they belong to. Fig. 1 shows an example of each approach on the amino-acid dataset.

The *tabular approach* uses a table or a matrix-derived view (Fig. 1, top left). It relies on the ability of the human vision to distinguish horizontal and vertical lines, better than other lines [3]. The elements are shown in rows and the sets in columns, or *vice versa*.

The *graph-based approach* displays elements and sets as nodes, and the memberships are shown as edges linking elements to the sets they belong to. An example is the node-link diagram in Jigsaw [4], which relies on a bipartite graph. Many algorithms exist for arranging the nodes [5]; however, graphs are often difficult to read when representing overlapping sets (this is the case for the amino-acid dataset, Fig. 1, bottom left).

The *positional approach* shows the elements, and their positions indicate the sets they belong to. Three sub-approaches can be distinguished. (1) Euler and Venn diagrams represent elements by dots and sets by closed curves containing all the dots that correspond to the elements in the set. In biology, a Venn diagram is traditionally used for representing amino-acid properties (Fig. 1, bottom right). Euler and Venn diagrams become more and more complex when the number of sets increases above four. Additionally, it is difficult to generate automatically Euler / Venn diagrams, although not impossible [6]. A variant uses lines instead of closed curves [7]. (2) Map-based techniques represent geographic datasets as overlays on a map. (3) Scatterplot-based techniques represent only the elements (sets are not visible) and elements belonging to the same (or similar) sets are placed together. They compute the distance between each pair of elements, based on their set memberships, and then project the elements on a two-dimensional plot. A well-known example is Principle Component Analysis (PCA).

The *faceted approach* represents only the sets; the elements are not shown individually but aggregated data is shown. These techniques target huge datasets. An example is Radial Sets [8] (Fig. 1, top right). This technique represents sets as segments of a circle. Each segment includes a histogram indicating the number of elements in the set that belong to zero, one, two,... , *n* other sets. An arc joins each pair of segments, and the weight of the arc is proportional to the number of cooccurrence, *i.e.* the number of elements belonging to the two sets. Interactivity allows the user to select a group of elements, and to see how they spread over the various sets (in Fig. 1, top right, the elements in the aliphatic set have been selected).

3. Rainbow boxes

3.1. General principles

Rainbow boxes are a new technique for overlapping set visualization, inspired by the tabular approaches. It aims at visualizing relatively small datasets in detail.

Fig. 2 shows rainbow boxes displaying the amino-acid properties. In rainbow boxes, the elements are displayed in columns. Each set is displayed as a rectangular box covering all the columns corresponding to the elements in the set; the label of the set is shown

[1] Actually, overlapping set visualization is a symmetric problem. Thus, in this situation, one might consider the drugs as the sets and the properties as the elements as well. However, in the rest of the paper, we will consider the drugs as the elements, because the proposed visualization technique, rainbow boxes, works better with fewer elements than sets.