Information Processing Letters 115 (2015) 689–693

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Suppression distance computation for hierarchical clusterings

François Queyroi [a,b,*], Sergey Kirgizov [a,b]

[a] Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, France
[b] CNRS, UMR 7606, LIP6, F-75005, Paris, France

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form 14 April 2015
Accepted 15 April 2015
Available online 20 April 2015
Communicated by B. Doerr

Keywords:
Algorithms
Hierarchical partition
Clustering
Distance
Graphs

ABSTRACT

We discuss the computation of a distance between two hierarchical clusterings of the same set. It is defined as the minimum number of elements that have to be removed so the remaining clusterings are equal. The problem of distance computing was extensively studied for partitions. We prove it can be solved in polynomial time in the case of hierarchies as it gives birth to a class of perfect graphs. We also propose an algorithm based on recursively computing maximum assignments.

1. Introduction

Decomposing a set into patterns of interest is a central problem in data analysis. Evaluating the distance between decompositions is an important task in this context as it allows to study the behaviour of clustering algorithms or study the evolution of a set of patterns over time. The situation where the detected patterns do not overlap is called *partitions*. Measures based on edit distance [3] or on mutual information [6] can be used to assess the distance between those objects. The first corresponds to the minimum number of elements that need to be moved from one group to another for the two partitions to be equal (called *transfer distance* in [7]). It was used for practical applications in bioinformatics [10]. Similar definitions can also applied to different kind of decompositions *e.g.* with overlapping groups (called *set covers*).

This work focuses on *hierarchical clusterings* (also called hierarchies) in which each group can be recursively decomposed into smaller groups. The problem of distance definition between hierarchies is of interest as they can be used to represent and study a system (such as *complex networks* [8]) at different scales. Comparing hierarchical clusterings is related to the comparison of phylogenetic trees [9] in biology although those objects have typically more constraints than the decompositions studied here.

We investigate the problem of finding the minimum number of elements to be removed so that the remaining hierarchical clusterings are equal or, equivalently, the size of smallest subset of elements for which the decompositions "disagree". After having define the core concepts (Section 2), we will provide two alternative proofs of the main claim (Sections 3 and 4). The first links the problem to a class of perfect graphs (generalizing the results of [3]) since the difference between hierarchies can be encoded into a graph (called the *difference graph*) with specific characteristics. The second provides a polynomial algorithm to compute the distance between hierarchical clusterings. Both approaches are based on similar observations (Lemmas 2 and 3). Section 5 provides concluding remarks and directions for future work.

* Corresponding author.
E-mail addresses: francois.queyroi@parisgeo.cnrs.fr (F. Queyroi), sergey.kirgizov@u-bourgogne.fr (S. Kirgizov).

http://dx.doi.org/10.1016/j.ipl.2015.04.007
0020-0190/

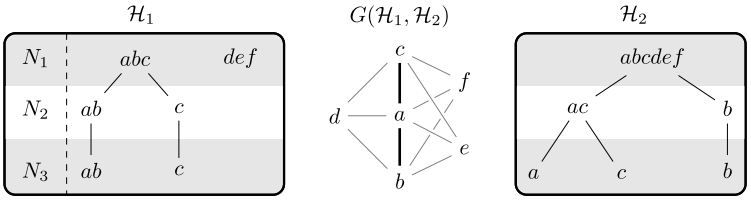

Fig. 1. Example of two hierarchies $\mathcal{H}_1, \mathcal{H}_2$ of a set $S = \{a, b, c, d, e, f\}$ and their difference graph $G(\mathcal{H}_1, \mathcal{H}_2)$. The levels of $\mathcal{H}_1$ are $N_1(\mathcal{H}_1) = \{\{a, b, c\}, \{d, e, f\}\}$ and $N_2(\mathcal{H}_1) = \{\{a, b\}, \{c\}\}$. We have $d_S(\mathcal{H}_1, \mathcal{H}_2) = 3$ with the suppression set $S' = \{a, b, c\}$.

2. Definitions

We assume we have a set S of elements of finite cardinality. A *hierarchy* $\mathcal{H} = (H_1, H_2, \ldots, H_k)$ is a finite multiset of non-empty subsets of S such that if there exist two groups $H_1, H_2 \in \mathcal{H}$ such that if $H_1 \cap H_2 \neq \emptyset$ then either $H_1 \subseteq H_2$ or $H_2 \subseteq H_1$. The relation of inclusion between the sets defines a partial ordered set. It can be represented in a forest fashion, the roots of each tree being the sets that are not include in any other group.

Let $N_i(\mathcal{H})$ denotes the i-th level of $\mathcal{H}$ *i.e.* the groups sitting at depth i in this forest. Notice it is still well defined if $\mathcal{H}$ contains repeated groups. A level $N_i(\mathcal{H})$ is a partition since it does not contain overlapping sets. The *depth* of a hierarchy $d(\mathcal{H})$ is the maximum depth of its groups. We define as $\mathcal{H}[S']$ the *sub-hierarchy induced by* $S' \subseteq S$ as the non-empty sets of $\{S' \cap H_i\}_{1 \leq i \leq k}$. It is the hierarchical clustering of S' obtained after the removal of every elements of $\{S \setminus S'\}$ in each group of $\mathcal{H}$ (discarding empty sets).

Definition 1 *(Suppression distance).* Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hierarchies of S. The suppression distance d_s is defined as

$$d_s(\mathcal{H}_1, \mathcal{H}_2) = \min_{S' \subseteq S} \{|S'| : \mathcal{H}_1[S \setminus S'] = \mathcal{H}_2[S \setminus S']\}$$

A set S' such that $\mathcal{H}_1[S \setminus S'] = \mathcal{H}_2[S \setminus S']$ is called a *suppression set*.

Theorem 1. *The function d_s is a metric.*

Proof. The non-negativity, identity and symmetry properties are straightforward for d_s. Moreover, this distance respects the triangular inequality. Consider three hierarchies $\mathcal{H}_1, \mathcal{H}_2$ and $\mathcal{H}_3$. Let $S_{ij} \subseteq S$ be a minimum suppression set for $(\mathcal{H}_i, \mathcal{H}_j)$. Since $S_{12} \cup S_{23}$ is also a suppression set for $(\mathcal{H}_1, \mathcal{H}_3)$, we have:

$$|S_{13}| \leq |S_{12} \cup S_{23}| \leq |S_{12}| + |S_{23}|$$

$$d_s(\mathcal{H}_1, \mathcal{H}_3) \leq d_s(\mathcal{H}_1, \mathcal{H}_2) + d_s(\mathcal{H}_2, \mathcal{H}_3) \qquad \square$$

3. Existence of a polynomial-time solution

We give here a non-constructive proof for the existence of a polynomial time algorithm. It generalizes the results of Gusfield [3] on the equivalence between this problem and the minimum vertex cover problem on perfect graphs. The difference between hierarchies can be encoded in a *difference graph* (Definition 2). Finding a suppression set for two hierarchies is equivalent to find a minimum vertex cover in this graph (Theorem 2). Since, this graph is perfect [5] (Theorem 3), it exists a polynomial time algorithm to solve this problem.

Definition 2 *(Difference graph).* Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hierarchies of a set S. We call $G(\mathcal{H}_1, \mathcal{H}_2) = (S, E)$ the *difference graph* of $(\mathcal{H}_1, \mathcal{H}_2)$[1] with

$$E = \{(s_1, s_2) \in S^2 : |\mathcal{H}_1[\{s_1, s_2\}]| \neq |\mathcal{H}_2[\{s_1, s_2\}]|\}$$

This graph can contain self-loops.

Two elements of S are connected iff they do not appear in the same number of groups together in both hierarchies. An example of hierarchies and their difference graph can be found in Fig. 1.

Lemma 1. *Given $G = (S, E)$ the difference graph of $(\mathcal{H}_1, \mathcal{H}_2)$ and $S' \subseteq S$, the induced subgraph $G[S']$ is the difference graph of $(\mathcal{H}_1[S'], \mathcal{H}_2[S'])$.*

Proof. Let $G' = G(\mathcal{H}_1[S'], \mathcal{H}_2[S'])$. First, notice that $V(G') = V(G[S'])$ by definition. Second, we have $E(G') = E(G[S'])$. Indeed, for $i \in \{1, 2\}$, the number of groups where $\{s_1, s_2\} \in S'^2$ appear together is equal in $\mathcal{H}_i$ and $\mathcal{H}_i[S']$ by definition of induced hierarchy. Therefore, we have $E(G') = \{(s_1, s_2) \in S'^2, |\mathcal{H}_1[\{s_1, s_2\}]| \neq |\mathcal{H}_2[\{s_1, s_2\}]|\}$ which is also equal to $E(G[S'])$ by definition of induced subgraph. $\square$

Theorem 2. *$d_s(\mathcal{H}_1, \mathcal{H}_2)$ is equal to the size of the minimum vertex cover of $G(\mathcal{H}_1, \mathcal{H}_2)$.*

Proof. Let $G = G(\mathcal{H}_1, \mathcal{H}_2)$. We show first that $E(G) = \emptyset \Leftrightarrow \mathcal{H}_1 = \mathcal{H}_2$.

1. $(\mathcal{H}_1 = \mathcal{H}_2) \Rightarrow (E(G) = \emptyset)$ by definition of difference graph.
2. $(E(G) = \emptyset) \Rightarrow (\mathcal{H}_1 = \mathcal{H}_2)$
   (a) $d(\mathcal{H}_1) = d(\mathcal{H}_2) = d$ since G contains no self-loops by hypothesis. Every $s \in S$ belongs to the same number of sets in both hierarchies and $d(\mathcal{H}) = \max_{s \in S} |\mathcal{H}[\{s\}]|$.
   (b) $G = \bigcup_{i=1}^{d} G(N_i(\mathcal{H}_1), N_i(\mathcal{H}_2))$ since all elements in S belong to at most one group at a given level by definition of hierarchy. Indeed, let $(a, b) \in S^2$

[1] To simplify notations, G will sometimes be used instead of $G(\mathcal{H}_1, \mathcal{H}_2)$.