Minimum Spanning Tree Problem with Label Selection and Its Application to Mathematical OCR

Akio Fujiyoshi
Ibaraki University
fujiyosi@mx.ibaraki.ac.jp

Masakazu Suzuki
Kyushu University
suzuki@math.kyushu-u.ac.jp

Abstract

In this paper, we study the minimum spanning tree problem for vertex-labeled graphs, where the weights of edges may vary depending on the selection of labels of vertices at both ends. The problem is especially important as the application to mathematical OCR.

It is shown that the problem is NP-hard even on directed acyclic graphs (DAGs). However, there exists a linear-time algorithm for graphs of small tree-width. The relation to the generalized minimum spanning tree problem is discussed.

1 Introduction

The minimum spanning tree problem is one of the most famous combinatorial problems in computer science. Fast algorithms to solve the problem are well-known. In this paper, we study a generalization of the problem for vertex-labeled graphs, where the weight of edges may vary depending on the selection of labels of vertices at both ends.

An instance of the problem is illustrated in Fig. 1 (a): A finite set of symbols for vertex-labeling is $\Sigma = \{a, b, c, d\}$; Vertices are indicated by dotted rectangles; Each vertex has at least one candidates of labels represented by circled symbols; Each weighted edge connects labels that belong to different vertices; Some pairs of labels may not be connected. For this instance of the problem, the minimum spanning tree is illustrated in Fig. 1 (b): Exactly one label is selected from candidates for each vertex; The graph induced by selected candidates of labels and selected edges is a spanning tree. We also introduce the notion of a base graph. The corresponding base graph is illustrated in Fig. 1 (c). We say two vertices are connected if some candidates of two vertices are connected.

For the development of mathematical OCR [1], the problem is especially important. As shown in Fig. 2 (a) and (b), a mathematical OCR system constructs a DAG that expresses possible adjacency connections of bounding boxes from a scanned image. At this moment, several character recognition candidates may remain for each bounding box. Each edge is weighted by the positional relation and co-occurrence of character recognition candidates. In order to output a better recognition result as shown in Fig. 2 (c), the system wants to find the minimum spanning tree from the DAG not only by selecting character recognition candidates for bounding boxes but also by determining adjacency connections of bounding boxes.

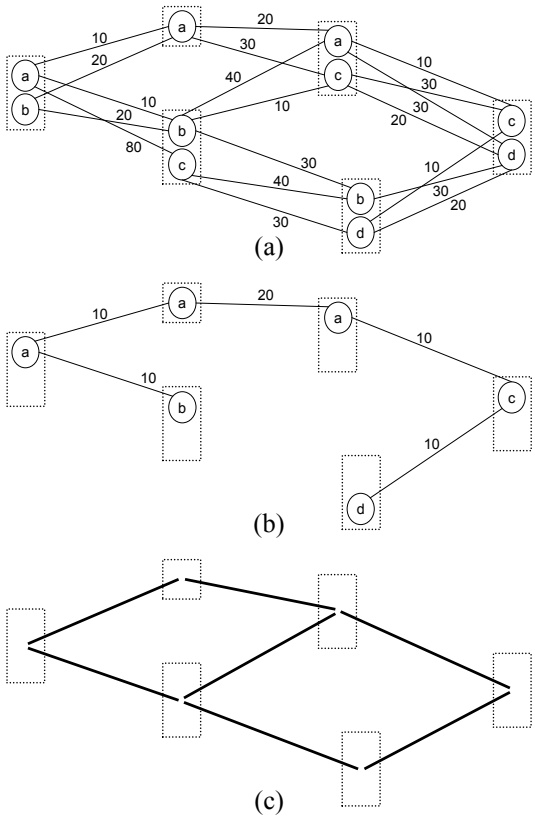

Figure 1: (a) an instance of the minimum spanning tree problem with label selection, (b) the minimum spanning tree, and (c) the base graph.

2 NP-Hardness

To show the NP-hardness, we reduce the Boolean satisfiability problem (SAT) to this problem.

Theorem 1 *The minimum spanning tree problem with label selection is NP-hard for directed graphs, and the problem is still NP-hard even on DAGs.*

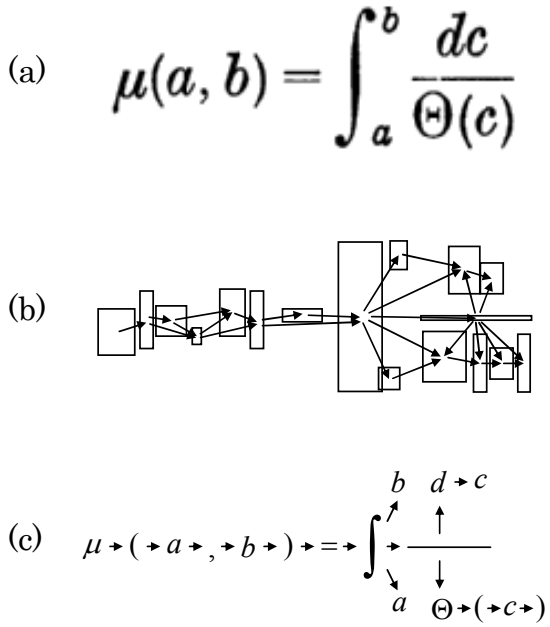

Figure 2: (a) a scanned image, (b) a DAG expressing possible adjacency connections of bounding boxes, (c) a recognition result.

Sketch of proof. Given a CNF formula, we can construct a DAG such that the DAG has a spanning tree if and only if the formula has a truth assignment. For a CNF formula $(x_1 \vee \bar{x_2} \vee x_5) \wedge (x_2 \vee x_3 \vee \bar{x_4}) \wedge (x_3 \vee \bar{x_4} \vee \bar{x_5})$, the corresponding DAG is illustrated in Fig. 3. □

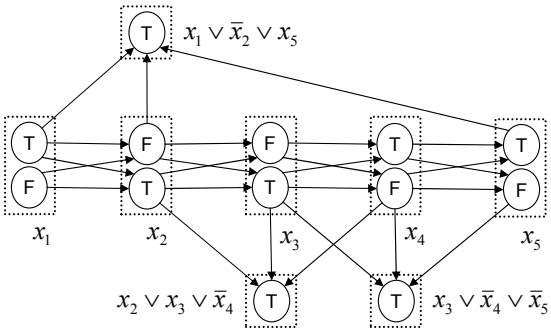

Figure 3: the DAG corresponding to the formula.

If we put weights on edges properly, the following corollary is obtained.

Corollary 1 *The minimum spanning tree problem with label selection is NP-hard for undirected graphs.*

3 Linear-Time Algorithm

We have implemented a linear-time algorithm for graphs whose base graph has tree-width at most 2. For the reasons of space, we only present the main idea of the algorithm as follows:

- When two graphs are connected in series, the minimum spanning tree is obtained by simply connecting the two minimum spanning trees of original graphs.
- When two graphs are connected in parallel, the minimum spanning tree is obtained by connecting the minimum spanning tree of one original graph and the minimum two disconnected spanning subtrees of the other graph.

We plans to implement a liner-time algorithm for graphs whose base graph has tree-width 3 or 4.

4 Relation to the Generalized Minimum Spanning Tree Problem

The minimum spanning tree problem with label selection is closely related to the generalized minimum spanning tree problem (GMSTP) [2, 3].

The following results are known for GMSTP [3]:

- GMSTP is NP-hard, and the problem is still NP-hard even on trees.
- If the number of clusters is fixed, then GMSTP can be solved in polynomial-time with respect to the number of vertices.

The result of this paper can be translated into the following new results for GMSTP:

- GMSTP is still NP-hard even if the size of each cluster is at most 2.
- If the size of each cluster is fixed, and the tree-width of the base graph is small, then GMSTP can be solved in linear-time with respect to the number of vertices.

References

[1] Eto, Y., Suzuki, M.: Mathematical formula recognition using virtual link network. In: Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR 2001). (2001) 430–437

[2] Myung, Y.S., Lee, C.H., Tcha, D.W.: On the generalized minimum spanning tree problem. Networks **26**(4) (1995) 231–241

[3] Pop, P.C.: The generalized minimum spanning tree problem. PhD thesis, Twente University Press, `http://doc.utwente.nl/38643/` (2002)