Information Processing Letters 115 (2015) 965–968

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Google Scholar makes it hard – the complexity of organizing one's publications

Hans L. Bodlaender, Marc van Kreveld *

Department of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TB, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 14 October 2014
Received in revised form 2 June 2015
Accepted 3 July 2015
Available online 9 July 2015
Communicated by A. Muscholl

Keywords:
Computational complexity
NP-completeness
Reduction
3-partition
Google Scholar

ABSTRACT

With Google Scholar, scientists can maintain their publications on personal profile pages, while the citations to these works are automatically collected and counted. Maintenance of publications is done manually by the researcher herself, and involves deleting erroneous ones, merging ones that are the same but which were not recognized as the same, adding forgotten co-authors, and correcting titles of papers and venues. The publications are presented on pages with 20 or 100 papers in the web page interface from 2012–2014. (Since mid 2014, Google Scholar's profile pages allow any number of papers on a single page.) The interface does not allow a scientist to merge two versions of a paper if they appear on different pages. This not only implies that a scientist who wants to merge certain subsets of publications will sometimes be unable to do so, but also, we show in this note that the decision problem to determine if it is possible to merge given subsets of papers is NP-complete.

1. Introduction

Most researchers in computer science will be familiar with Google Scholar and its abilities to maintain publications and their citations. Each researcher has his/her own profile which is shown as a web page with a list of publications. Google Scholar determines the number of citations to each publication and by default, lists them in this order on web pages for that researcher. Since the collection of the data is automated, it will contain various mistakes, many of which are caused by other scientists who fail to give the title or other essential information on a paper correctly. As a consequence, a single paper may have various versions in the list, with a slightly different title or publication venue, or with co-authors missing (see Fig. 1). Google Scholar offers researchers the possibility to correct these mistakes on their own profile page, for example by allowing them to merge two paper versions into one. This creates one version with the citations of the original versions summed up. Of course, one could also delete the erroneous version, but this may cost some citations, which on its turn can influence the ever-important H-index and other summary statistics that Google Scholar maintains.

Google Scholar by default shows the publication list on pages with 20 papers. It is possible to change this number to 100. Since mid 2014, a change in the interface makes it possible to get all publications on a single page. In this note we assume the interface in use from 2012 until mid 2014, when this was not possible and the maximum was 100 on a single page. To merge two papers, both should be selected on the web page, after which the merge action can be executed. However, selection of two papers is possible only if they appear on the same page, and therefore, merging can be done only if the two papers appear in the same group of 100 papers. For example, if one paper is the 103rd by citation count and another paper is the

* Corresponding author. Tel.: +31 30 2534119.
E-mail addresses: h.l.bodlaender@uu.nl (H.L. Bodlaender), m.j.vankreveld@uu.nl (M. van Kreveld).

http://dx.doi.org/10.1016/j.ipl.2015.07.003
0020-0190/

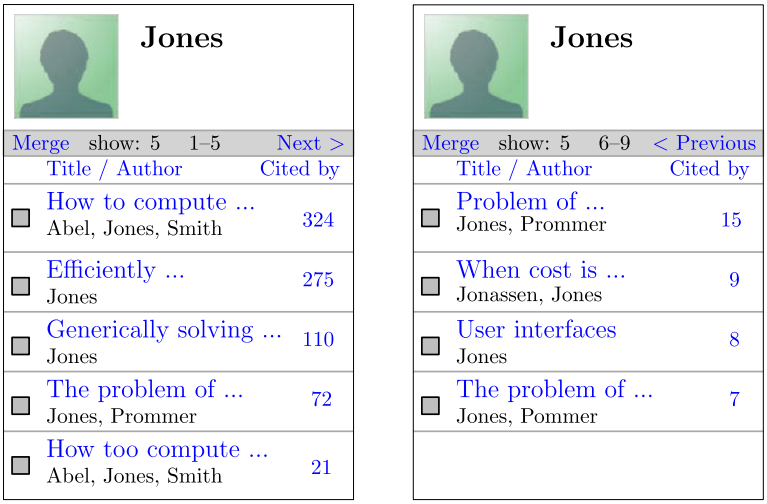

Fig. 1. Two pages for the author Jones with six different papers occurring as nine versions of papers. Jones must perform three merges to correct the data, and the order is important.

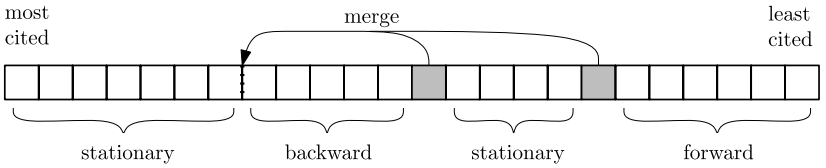

Fig. 2. A merge and the changes of positions of all versions of papers in the sorted list.

187th, then they can be merged, but if one is the 97th and the other is the 105th, then they cannot be merged.

The order in which papers are merged is important. If there are two pairs of papers to be merged, for example at positions 4 and 12, and at positions 101 and 107, then merging the first pair first will move the positions of the latter pair to 100 and 106, putting them on different pages. But merging the second pair first still allows the merging of the first pair. Notice that the position of a paper can change both forward and backward due to a merge, see Fig. 2.

Besides the problem that desired merges sometimes cannot be done, the computational problem of deciding whether a sequence of merges exists (and therefore, finding the correct order) is computationally intractable. This implies that a polynomial-time algorithm to produce the sequence of merges is unlikely to exist.

2. A proof of intractability

To prove intractability, or, NP-completeness of the problem, we will formalize it first. Let n be the total number of versions of papers initially in a problem instance, and let p be the page size. Let the paper versions be $v_1, \ldots, v_n$, and assume that paper version v_i is cited $c(v_i)$ times. A problem instance consists of the sequence $c(v_1), \ldots, c(v_n)$, and a partition of $1, \ldots, n$ into subsets where two or more versions in the same subset indicates that they are different versions of the same paper, and therefore, they are to be merged into one. The Google Scholar Merge Problem is the problem of deciding whether for every subset, all of its versions can be merged. When two versions are merged, they appear as one new version and their citations are added. After each merge, the new set of versions appears in sorted order on citation count. When citation counts are the same, the papers will appear in some other well-defined order, but this will be irrelevant for the intractability proof and we will ignore this issue.

Theorem 1. *The Google Scholar Merge Problem is NP-complete.*

Proof. First, we will verify that the Google Scholar Merge Problem is in NP. This is easy: a suggested merge order can easily be checked in quadratic time or less.

Second, we use another NP-complete or NP-hard problem and provide a reduction to our problem, namely 3-PARTITION [2]. A 3-PARTITION instance consists of a set $a_1, \ldots, a_{3m}$ of positive integers and an integer B, and asks whether we can partition $a_1, \ldots, a_{3m}$ in m subsets of 3 integers that each sum up to B. The integers $a_1, \ldots, a_{3m}$ are all strictly between $B/4$ and $B/2$, which ensures that any subset that sums to B consists of exactly three integers.

We describe the reduction from 3-PARTITION to the Google Scholar Merge Problem. First, we double all a_i and B to ensure that they are even. With slight abuse of notation we continue to use the notation $a_1, \ldots, a_{3m}$ and B for these doubled values.

We set the page size to $3m$. Let D be some large, even integer; we can use $D = 3mB$.

Our instance consists of one paper P with many versions and many papers with one version, see Fig. 3 for a schematic depiction.