Counting word avoiding factors*

Manuel Baena-García, Rafael Morales-Bueno, José M. Carmona-Cejudo
mbaena@lcc.uma.es, morales@lcc.uma.es, jmcarmona@lcc.uma.es
Lenguajes y Ciencias de la Computación
Universidad de Málaga
Spain

Gladys Castillo†
gladys@ua.pt
Department of Mathematics, CEOC
University of Aveiro
Portugal

Abstract

Combinatorics is a branch of mathematics strongly related to computer science, with a very prominent role in school curricula. In order to develop the "combinatorial thinking" of students, it is important to provide them with detailed demonstrations of fundamental theorems and results in this area, so that they can get a "feel" of combinatorics. With this in mind, we provide an algorithmic solution for a given combinatorial problem, the avoiding factors problem, with a detailed demonstration, at the level of an undergraduate student. A website has been published with the aim of letting students experiment with an implementation of our solution, enabling them to verify their own work.

1 Introduction

In this work we deal with *combinatorics*, a branch of mathematics that studies discrete objects, which is considered to be foundational in computer science [7]. It can be defined as the *study of ways to list and arrange elements of discrete sets according to specified rules* [4]. *Combinatorics in words* is a subfield of combinatorics applied to words and formal languages, and it's of great importance for theoretical computer science. Furthermore, it has led to basic algorithms for text processing and bioinformatics, among others.

*This work has been partially supported by the SESAME project, number TIN2008-06582-C03-03, of the MICINN, Spain.

†Supported by FCT, POCI (EC fund FEDER) through CEOC

Due to the importance of this area for computer science, it is being given an increasingly prominent role in school curricula in recent years. From the teacher's point of view, it is desirable to provide strategies that help students not only to develop their combinatorial thinking, but also to elaborate strategies to let them verify their solutions to the proposed combinatorics problems. One way to develop their combinatorial understanding is to offer them examples of solutions to several problems, following a 'learn-by-example'-approach. Through the study of these examples, the student will become familiar with the elementary steps that make up complex formal demonstrations in this field.

In this work we propose a relatively well-known problem, and give a new solution for it, developed by the authors. The demonstration is elaborated in a very detailed way, making it possible for students with little background to follow and understand it, and acquiring abilities to follow and elaborate similar solutions.

Additionally, we provide a computer implementation of the solution, with the aim of assisting them to verify their solution with the help of a computer. This implementation is published under the GNU GPL license, meaning that they are free to modify and distribute it.

2 Informal Definition of the problem

Given a factor (a short string), a typical task is to find its occurrence in a much longer string (a text) using a string search algorithm. Several combinatorics problems can be formulated in relation to this task, such as counting the number of words on a finite alphabet so that a finite set of factors is avoided. The solution to this problem lets us estimate the spatial growth of the data structures involved, and can be used to adjust heuristics for avoiding memory overflow. This relatively simple problem has been widely studied, and quite a few solutions have appeared in the literature.The best known solution can be found in the work of Guibas and Odlyzko [6].

Let's give a simple example to illustrate the problem, to be found in Tanya Khovanonova's Web page[1]

Proposition 1. *The number of words of length n in the alphabet $\{1, 2 \ldots, d\}$ avoiding words with "12" as substring is equal to $a(n)$, a recursive sequence with initial terms $a(0) = 1$, $a(1) = d$ and recurrence relation*

$$a(n) = d \times a(n-1) - a(n-2) \tag{1}$$

Proof. Let us denote $b(n)$ (correspondingly $c(n)$) to the number of words in this sequence of length n that end (correspondingly that do not end) in 1. Hence $a(n) = b(n) + c(n)$.

You can see that $b(n) = a(n-1)$ since we can obtain a valid word ending in 1 by attaching 1 to any valid word.

We can obtain a word not ending in 1 by attaching any digit other than 1 and 2 to any valid word, or by attaching 2 to any valid word not ending in 1. Hence $c(n) = (d-2) \times a(n-1) + c(n-1)$. Now we can replace $c(n-1)$ as $a(n-1)-b(n-1)$, which give us $c(n) = (d-2)\times a(n-1)-a(n-1)-b(n-1) = (d-1) \times a(n-1) - b(n-1)$.

Adding b and c we obtain $a(n) = (d-1) \times a(n-1) - a(n-2)$. □

[1] `http://www.tanyakhovanova.com/RecursiveSequences/RecursiveSequences.html`

The sequences generated by this equation can be found in [10]. The first terms of this recurrence for $d = 3$ are 1, 3, 8, 21, 55, 144, 377, 987, ... (A001906)

Such a solution shows, step by step, the reasoning processes involved in a typical demonstration of a theorem on string combinatorics, and can help novel students to become familiar with this sort of demonstrations, as well as with some useful results in this field.

3 The avoiding factor problem

Notations and Basic Concepts

Let Σ be a finite set of characters (or symbols) called *alphabet*. A *string* (or *word*) over Σ is a sequence of characters of Σ. The length of a string w is the number of its characters and is denoted $|w|$. The characters of w are indexed from 1 to $|w|$, the i-th character of w is denoted w_i. A string can also have length 0. The empty string is denoted ϵ ($|\epsilon| = 0$). A *substring*, *prefix* or *suffix* of a string is a subset of the characters in w, where the order of the elements is preserved. Let $w = w_1 \dots w_{|n|}$ be a string of length n. A *substring* of w is a string $e = w_{i+1} \dots w_{i+|e|}$, where $i \geq 0$ and $i + |e| \leq n$. Otherwise, we say that e *avoid as substring* w and denote it as $e \not\sqsubseteq w$. A *prefix* of w is a string $p = w_1 \dots w_{|p|}$, where $|p| \leq n$. If $0 < |p| < n$ we call p a *proper prefix* of a word w, that is, a proper prefix is not equal to the string itself and is not empty. A suffix of w is a string $s = w_{n-|s|+1} \dots w_{|w|}$, where $|s| \leq n$. Given two strings x and y over Σ, the *concatenation* of x and y is the string xy obtained by appending y at the end of x. Σ^* is the set of all words over Σ. Σ^n is the set of all words over Σ that have length n. Σ^+ is the set of all words over Σ that have length ≥ 1. By language misuse σ will be used as both symbol and word.

The Problem

Consider the problem of determining the number of words in Σ^n that avoid a set of factors. Let $E \subset \Sigma \cdot \Sigma^+$ be the set of avoided factors.

Definition 2. *Given an alphabet Σ and a subset of non overlapped avoidable factors E, the* avoidance factor problem $a_E(n)$ *is defined as the number of words of length n over the alphabet Σ that avoid as substring every word in E, that is*

$$\begin{aligned} a_E : \ \mathbb{N} &\to \mathbb{N} \\ n &\to a_E(n) = |W_n| \end{aligned} \tag{2}$$

$$W_n = \{w \in \Sigma^n \ : \ \forall e \ [e \in E \to e \not\sqsubseteq w]\} \tag{3}$$

Remark 3. *We call E the* set of avoiding factors. *Non overlapped means that the words in the set E have the following property: $\forall w \forall y \ [w, y \in E \to (w \not\sqsubseteq y \ \wedge \ y \not\sqsubseteq w)]$*

Remark 4. *We call W_n the* set of valid words *of length n*

Example 5. *Suppose $\Sigma = \{0, 1\}$ and $E = \{00, 010\}$. The values of $a_E(n)$ for $n = 0, 1, 2, 3$ are $a_E(0) = 0$, $a_E(1) = 2$, $a_E(2) = 3$ and $a_E(3) = 4$.*

Figure 1 depicts a partial tree structure to represent the words in $\{W_n : n \in \mathbb{N}\}$ of the Example 5 where each level d represents the set W_d.

$a_E(0) = 0$	
Σ^0	W_0
$\emptyset$	$\emptyset$

$a_E(1) = 2$	
Σ^1	W_1
0	0
1	1

$a_E(2) = 3$	
Σ^2	W_2
00	-
01	01
10	10
11	11

$a_E(3) = 4$	
Σ^3	W_3
000	-
001	-
010	-
011	011
100	-
101	101
110	110
111	111

Table 1: Some values for $a_E(n)$

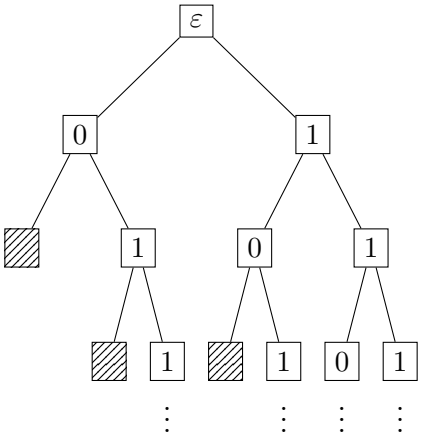

Figure 1: Tree for $E = \{00, 010\}$

4 A solution for the avoidance factor problem

In this section, we present our solution for the avoidance factor problem. It should be noted that other solutions have been previously proposed in the literature [6]. Similar to ours, but a more straightforward solution can be obtained using the Knuth-Morris-Pratt factor searching algorithm in its extended form, generating a finite state automaton whose rejecting paths represent those strings that avoid the factors. The length of these paths can be counted by length using the standard transfer matrix methodology. The recurrences used in our solution would represent the allowed transitions from and between states.

Nevertheless, such a demonstration would involve mathematical techniques that, generally, cannot be assumed to be known by students. That is why we propose our solution, more appropiated to the level of an undergraduate computer science student, although somehow more indirect.

Before we start the formal exposition that follows, we want to give the main ideas of the proposed

solution in order to allow the reader to better understand the key concepts and methods involved in the derived results. Given a set of elements, it is often useful to partition them into a number of separate, non overlapping (disjoint) sets. The rationale of our solution is to choose a suitable partition of the set of valid words so that the problem of counting the number of valid words that avoid a set of factors can be reduced to the problem of determining the cardinality of each disjoint partition. For this purpose, the proof is divided into two steps. In the first step we will show that the minimal set of proper prefixes of words in the set of avoiding factors let us find a good partition of the set of valid words. We will prove that the set of valid words can be decomposed into the set of valid words ending in a proper prefix of avoiding factors and the set of valid words not ending in them. In the second step we introduce another kind of set, called *linking sets*, and will use it to determine the cardinality of the valid-prefix by using concatenation over sets. The general solution is described as a recurrence relation system.

4.1 Partitioning the set of valid words into disjoint sets

Let W_n be a set of valid words of length n, E a set of avoiding factors and $p(E)$ the minimal set of proper prefixes of words in E.

Definition 6. *A* ϵ-valid-prefix set $X_n^\epsilon \subset W_n$ *is the subset of valid words in* W_n *that not ending in a proper prefix of a word in* E*, that is,*

$$X_n^\epsilon = \left\{ w \;\middle|\; w \in W_n \wedge \forall x \forall t \in \Sigma^+ \left[w = xt \rightarrow t \notin p(E) \right] \right\} \tag{4}$$

Definition 7. *A* t-valid-prefix set $X_n^t \subset W_n$ *is the subset of valid words in* W_n *ending in* t *with* t *being the bigger suffix that is a proper prefix in* $p(E)$*, that is,*

$$\begin{aligned} X_n^t = \Big\{ w \;\Big|\; & w \in W_n \;\wedge\; \exists x : w = xt \;\wedge\; t \in p(E) \wedge \\ & \forall x \forall y \in \Sigma^+ \left[w = xyt \rightarrow yt \notin p(E) \right] \Big\} \end{aligned} \tag{5}$$

Definition 8. *A* valid-prefix set X_n^T *is the union of the* t*-valid-prefix sets, that is,* $X_n^T = \bigcup_{t \in p(E)} X_n^t$

Example 9. *Let* $E = \{00, 010\}$ *be the set of avoiding factors in Example 5. Then for* $n = 3$ *we have* $W_3 = \{011, 101, 110, 111\}$ *(see Table 1). The set of proper prefixes* $p(E) = \{0, 01\}$ *defines two* t*-valid-prefix sets* X_3^0 *and* X_3^{01}*. Then due to the above definitions* $X_3^\epsilon = \{011, 111\}$*,* $X_3^0 = \{110\}$*,* $X_3^{01} = \{101\}$ *and* $X_3^T = \bigcup_{t \in p(E)} = \{110, 101\}$*. Note that* $X_3^0 \cap X_3^{01} = \emptyset$ *and* $X_3^\epsilon \cap X_3^T = \emptyset$*. As a result* W_3 *can be decomposed into two disjoint (on the union) sets, that is,* $W_3 = X_3^\epsilon \cup X_3^T$ *with* $X_3^\epsilon \cap X_3^T = \emptyset$*.*

Lemma 10. *The union of the sets* X_n^ϵ *and* X_n^T *is* W_n*, that is,* $W_n = X_n^\epsilon \cup X_n^T$

Proof. By the definition of union of sets we must prove that for every w, $w \in W_n$ *if and only if*, $w \in X_n^\epsilon \vee w \in X_n^T$. To do this we must first prove that for every w, if $w \in W_n$ then $w \in X_n^\epsilon \vee w \in X_n^T$. We proceed by contradiction. Suppose that there exists $w \in W_n$ such that $w \notin X_n^\epsilon \;\wedge\; w \notin X_n^T$. We consider two cases:

Case 1:if $w \notin X_n^\epsilon$, then by Definition 6 $\exists\, t \in \Sigma^+$ such that if $w = xt$ then $t \in p(E)$. By Definition 7 and Definition 8 it follows that $w \in X_n^T$, which is a contradiction.

Case 2: if $w \notin X_n^t$, then by Definition 7 it follows that $t \notin p(E) \ \vee \ \exists x \exists y \in \Sigma^+ : \ (w = xyt \ \wedge \ yt \in p(E))$. This implies $w \in X_n^\epsilon \vee w \in X_n^u$ with $u \neq t$. This is also a contradiction

In the other direction, suppose that for every w, $w \in X_n^\epsilon \vee w \in X_n^T$. Since we know that $X_n^\epsilon \subseteq W_n$, $\forall t \ X_n^t \subseteq W_n$, then $X_n^T \subseteq W_n$; hence $X_n^\epsilon \cup X_n^T \subseteq W_n$. As a result $w \in W_n$. □

The above lemma can be interpreted as "The union of valid words ending in a proper prefix of E and the valid words not ending in a proper prefix of E define the set of every possible valid words".

Lemma 11. *The intersection of the sets $X_n^t \cap X_n^u$ $(t \neq u)$ and the intersection of the sets $X_n^t \cap X_n^\epsilon$ is empty.*

Proof. This is obvious for X_n^ϵ. We obtain the contradiction $\forall x \forall t \in \Sigma^+ \ [w = xt \rightarrow t \notin p(E)]$ and $t \in p(E)$.

We study the intersection for X_n^t and X_n^u. We suppose $w \in X_n^t$ and $w \in X_n^u$. Then, $t = vu \vee u = vt$, with $v \neq \epsilon$. In both cases we have a contradiction for the condition in (5): $\forall x \forall y \in \Sigma^+ \ [w = xyt \rightarrow yt \notin p(E)]$ with $y = v$ □

Note: This is simple for words ending or not ending in a prefix of E. For the second case, intersection between sets t and u, each valid word belongs to the set determined by the words in E which have bigger overlapping.

After having found a suitable partition of the set of valid words we can use the cardinality of the partitions X_n^T and X_n^ϵ to count the total number of valid words avoiding the set of factors E.

Lemma 12. *$a_E(n)$ is equal to the sum of the cardinality of the set X_n^ϵ and the cardinality of the set X_n^T, that is.*

$$a_E(n) = |X_n^\epsilon| + |X_n^T| = |X_n^\epsilon| + \sum_t |X_n^t| \tag{6}$$

Proof. By the lemma 10 we know that the union of the sets X_n^ϵ and X_n^T is W_n. By the lemma 11 we know that their intersection is empty. Hence, the cardinal of $a_E(n)$ is the sum of the cardinals of the sets X_n^ϵ and X_n^T □

4.2 Determining the cardinality of the valid-prefix sets

To determine the cardinality of the valid prefix sets X_n^ϵ and X_n^T let us introduce the *linking sets* $C_\square^\square$ where the superindex can be a symbol or a word and the subindex is a word.

Definition 13. *Let $\sigma \in \Sigma$. The linking set C_ϵ^σ is a set with a symbol σ if this symbol is a proper prefix in E. Otherwise C_ϵ^σ is an empty set.*

$$C_\epsilon^\sigma = \left\{ \ \sigma \ \middle| \ \sigma \in \Sigma, \ \sigma \in p(E) \ \right\} \tag{7}$$

Definition 14. *Let u, and t be proper prefixes in E, with $u \neq t$. The linking set C_u^t is a set with one symbol σ if t is the bigger suffix for $u\sigma$ that is a proper prefix in E. Otherwise the set is empty. That is,*

$$\begin{aligned} C_u^t = \Big\{ \ \sigma \ \Big| \ & \sigma \in \Sigma, \ u \in p(E), \ t \in p(E), \ \exists s \colon \ u\sigma = st, \\ & \forall v \forall w \in \Sigma^+ \ [u\sigma = vwt \ \rightarrow \ wt \notin p(E)] \ \Big\} \end{aligned} \tag{8}$$

Definition 15. *Let t be a proper prefix in E. The linking set C_t^ϵ is the set of symbols σ such that any suffix of $t\sigma$ is not a prefix in E, that is,*

$$C_t^\epsilon = \{\ \sigma \mid \sigma \in \Sigma,\ t \in p(E),\ t\sigma \notin E,\ \forall x \forall y\ [t\sigma = xy\sigma\ \rightarrow\ y\sigma \notin p(E)]\ \} \tag{9}$$

Definition 16. *The linking set C_ϵ^ϵ is the set of symbols that are not a proper prefix in E, that is*

$$C_\epsilon^\epsilon = \{\ \sigma \mid \sigma \in \Sigma,\ \sigma \notin p(E)\ \} \tag{10}$$

Example 17. *Let $E = \{00, 010\}$ be the set of avoiding factors in Example 5 and $p(E) = \{0, 01\}$ the set of proper prefixes of Example 9. Then we have the linking sets $C_\epsilon^0 = \{0\}$, $C_0^\epsilon = \varnothing$, $C_0^0 = \varnothing$, $C_0^{01} = \{1\}$, $C_{01}^\epsilon = \{1\}$, $C_{01}^0 = \varnothing$, $C_{01}^{01} = \varnothing$ and $C_\epsilon^\epsilon = \{1\}$*

In the next lemmas we show that if we concatenate valid-prefix sets of length n with the linking sets we get valid-prefix sets of length $n + 1$.

Lemma 18. $\forall u \forall t\ [u, t \in p(E)\ \rightarrow\ X_n^u \cdot C_u^t \subseteq X_{n+1}^t]$.

Proof. We show that $\forall w \forall \sigma\ [w \in X_n^u\ \wedge\ \sigma \in C_u^t\ \rightarrow\ w\sigma \in X_{n+1}^t]$.

$$\begin{aligned}
&\text{From } w \in X_n^u \quad \equiv && w \in W_n,\ \exists x : w = xu,\ u \in p(E),\\
&&& \forall x \forall y \in \Sigma^+\ [w = xyu \rightarrow yu \notin p(E)]\\
&\text{and } \sigma \in C_u^t \quad \equiv && \sigma \in \Sigma,\ u \in p(E),\ t \in p(E),\ \exists s :\ u\sigma = st,\\
&&& \forall v \forall \dot{w} \in \Sigma^+\ [u\sigma = v\dot{w}t\ \rightarrow\ \dot{w}t \notin p(E)]\\
&\text{we will obtain } \Longrightarrow w\sigma \in X_{n+1}^t \quad \equiv &&\\
&&& w\sigma \in W_{n+1},\ \exists \dot{x} : w\sigma = \dot{x}t,\ t \in p(E),\\
&&& \forall \dot{x} \forall \dot{y} \in \Sigma^+\ [w\sigma = \dot{x}\dot{y}t \rightarrow \dot{y}t \notin p(E)]
\end{aligned}$$

We prove the implication for each term of the consequent.

- $\exists \dot{x} :\ w\sigma = \dot{x}t$. We know that $w = xu$ and $\exists s :\ u\sigma = st$, hence $w\sigma = xu\sigma = xst = \dot{x}t$ with $\dot{x} = xs$.
- $t \in p(E)$. By definition of C_u^t we know that this condition is true.
- $\forall \dot{x} \forall \dot{y} \in \Sigma^+\ [w\sigma = \dot{x}\dot{y}t \rightarrow \dot{y}t \notin p(E)]$. We know that

  $$\dot{w}t \notin p(E) \tag{11}$$

  and

  $$yv\dot{w}t \notin p(E) \tag{12}$$

  Furthermore, $w\sigma = xyv\dot{w}t$ and $w\sigma = \dot{x}\dot{y}t$. Since $\dot{y} \in \Sigma^+$ and $y, \dot{w} \in \Sigma^+$ the limit cases are

  $$\dot{y} = \dot{w}\ \rightarrow\ \dot{x} = xyv \tag{13}$$

  and

  $$\dot{y} = yv\dot{w}\ \rightarrow\ \dot{x} = x \tag{14}$$

  In case (13) we must consider (11) then $\dot{y}t \notin p(E)$. In case (14) we must consider (12), so $\dot{y}t \notin p(E)$. Hence, in every case $\dot{y}t \notin p(E)$.

- $w\sigma \in W_{n+1}$. We know that $w \in W_n$. By the implication defined previously, we know that $w\sigma$ does not have a suffix that is also a word in E. Hence $w\sigma \in W_{n+1}$.

□

Lemma 19. $X_n^{\epsilon} \cdot C_{\epsilon}^{\sigma} \subseteq X_{n+1}^{\sigma}$.

Proof. We prove that $\forall w \forall \sigma \; [w \in X_n^{\epsilon} \; \wedge \; \sigma \in C_{\epsilon}^{\sigma} \; \rightarrow \; w\sigma \in X_{n+1}^{\sigma}]$.

$$\begin{aligned}
&\text{From } w \in X_n^{\epsilon} \quad \equiv && w \in W_n, \\
& && \forall x \forall t \in \Sigma^+ \; [w = xt \; \rightarrow \; t \notin p(E)] \\
&\text{and } \sigma \in C_{\epsilon}^{\sigma} \quad \equiv && \sigma \in \Sigma, \; \sigma \in p(E) \\
&\text{we will obtain} \implies w\sigma \in X_{n+1}^{\sigma} \quad \equiv & & \\
& && w\sigma \in W_{n+1}, \; \exists \dot{x} : w\sigma = \dot{x}\sigma, \; \sigma \in p(E), \\
& && \forall \dot{x} \forall \dot{y} \in \Sigma^+ \; [w\sigma = \dot{x}\dot{y}\sigma \rightarrow \dot{y}\sigma \notin p(E)]
\end{aligned}$$

And prove the implication for each component of the consequent.

- $w\sigma = \dot{x}\sigma$ with $\dot{x} = w$.
- $\sigma \in p(E)$. By definition of C_{ϵ}^{σ}.
- $\forall \dot{x} \forall \dot{y} \in \Sigma^+ \; [w\sigma = \dot{x}\dot{y}\sigma \rightarrow \dot{y}\sigma \notin p(E)]$. We have that $\forall x \forall t \in \Sigma^+ \; [w = xt \rightarrow t \notin p(E)]$, that imply $w\sigma = \dot{x}\dot{y}\sigma \rightarrow \dot{y}\sigma \notin p(E)$ for $\dot{x} = x$ and $\dot{y} = t$.
- $w\sigma \in W_{n+1}$. We know that $w \in W_n$. By the above implication we know that $w\sigma$ does not have a prefix that is a word of E. Hence, $w\sigma$ is a valid word.

□

Lemma 20. $\forall t \; [t \in p(E) \; \rightarrow \; X_n^t \cdot C_t^{\epsilon} \subseteq X_{n+1}^{\epsilon}]$.

Proof. We analyze that $\forall w \forall \sigma \; [w \in X_n^t \; \wedge \; \sigma \in C_t^{\epsilon} \; \rightarrow \; w\sigma \in X_{n+1}^{\epsilon}]$.

$$\begin{aligned}
&\text{From } w \in X_n^t \quad \equiv && w \in W_n, \; \exists x : w = xt, \; t \in p(E), \\
& && \forall x \forall y \in \Sigma^+ \; [w = xyt \rightarrow yt \notin p(E)] \\
&\text{and } \sigma \in C_t^{\epsilon} \quad \equiv && \sigma \in \Sigma, \; t \in p(E), \\
& && \forall x \forall y \; [t\sigma = xy\sigma \; \rightarrow \; y\sigma \notin p(E)] \\
&\text{we can conclude} \implies w\sigma \in X_{n+1}^{\epsilon} \quad \equiv && w\sigma \in W_{n+1} \\
& && \forall \dot{x} \forall \dot{t} \in \Sigma^+ \; [w\sigma = \dot{x}\dot{t} \; \rightarrow \; \dot{t} \notin p(E)]
\end{aligned}$$

We prove each component of the consequent.

- $\forall \dot{x} \forall \dot{t} \in \Sigma^+ \; [w\sigma = \dot{x}\dot{t} \; \rightarrow \; \dot{t} \notin p(E)]$. Beginning from $\forall x \forall y \in \Sigma^+ \; [w = xyt \rightarrow yt \notin p(E)]$ can be implied that $\forall x \forall y \in \Sigma^+ \; [w\sigma = xyt\sigma \rightarrow yt\sigma \notin p(E)]$. Moreover, we know that $\forall x \forall y \; [t\sigma = xy\sigma \rightarrow y\sigma \notin p(E)]$. Hence, we have the implication for $\forall \dot{x} \forall \dot{t} \in \Sigma^+ \; [w\sigma = \dot{x}\dot{t} \rightarrow \dot{t} \notin p(E)]$.

- $w\sigma \in W_{n+1}$. We know that w is a valid word and by the above result we know that $w\sigma$ does not have a suffix which is a word in E. Hence, $w\sigma$ is a valid word.

□

Lemma 21. $X_n^\epsilon \cdot C_\epsilon^\epsilon \subseteq X_{n+1}^\epsilon$.

Proof. We study that $\forall w \forall \sigma\ [w \in X_n^\epsilon\ \wedge\ \sigma \in C_\epsilon^\epsilon\ \rightarrow\ w\sigma \in X_{n+1}^\epsilon]$.

$$\begin{aligned}
&\text{From } w \in X_n^\epsilon \quad \equiv && w \in W_n,\\
&&& \forall x \forall t \in \Sigma^+\ [w = xt\ \rightarrow\ t \notin p(E)]\\
&\text{and } \sigma \in C_\epsilon^\epsilon \quad \equiv && \sigma \in \Sigma,\ \sigma \notin p(E)\\
&\text{we can conclude } \Longrightarrow w\sigma \in X_{n+1}^\epsilon \quad \equiv && w\sigma \in W_{n+1},\\
&&& \forall \dot{x} \forall \dot{t} \in \Sigma^+\ [w\sigma = \dot{x}\dot{t}\ \rightarrow\ \dot{t} \notin p(E)]
\end{aligned}$$

We prove it for each term of the consequent:

- Let us see that $\forall \dot{x} \forall \dot{t} \in \Sigma^+\ [w\sigma = \dot{x}\dot{t}\ \rightarrow\ \dot{t} \notin p(E)]$.

  By beginning with $\forall x \forall t\ \in\ \Sigma^+\ [w = xt \rightarrow t \notin p(E)]$ we obtain that $\forall x \forall t \in \Sigma^+\ [w\sigma = xt\sigma \rightarrow t\sigma \notin p(E)$. Also, we know that $\sigma \notin p(E)$. Hence, we have that $\forall \dot{x} \forall \dot{t} \in \Sigma^+\ [w\sigma = x\dot{t} \rightarrow \dot{t} \notin p(E)]$.

- $w\sigma \in W_{n+1}$. We know that $w \in W_n$. Also we know that $w = xt \rightarrow t \notin p(E)$, we say, w does not have a suffix belonging to $p(E)$. Hence $w\sigma$ is a valid word.

□

Now, we will show a reverse version of the previous lemmas: if two words belong to consecutive valid-prefix sets and they are equal except for the last symbol, then that symbol is in the corresponding linking set.

Lemma 22. $w\sigma \in X_{n+1}^t \wedge w \in X_n^u \rightarrow \sigma \in C_u^t$.

Proof. We show that

$$\begin{aligned}
&\text{from } w\sigma \in X_{n+1}^t \quad \equiv && w\sigma \in W_{n+1},\ \exists x :\ w\sigma = xt,\ t \in p(E),\\
&&& \forall x \forall y \in \Sigma^+\ [w\sigma = xyt \rightarrow yt \notin p(E)]\\
&\text{and } w \in X_n^u \quad \equiv && w \in W_n,\ \exists y :\ w = yu,\ u \in p(E),\\
&&& \forall x \forall y \in \Sigma^+\ [w = xyu \rightarrow yu \notin p(E)]\\
&\text{we can conclude } \Longrightarrow \sigma \in C_u^t \quad \equiv &&\\
&&& \sigma \in \Sigma,\ u \in p(E),\ t \in p(E),\ \exists \dot{s} :\ u\sigma = \dot{s}t,\\
&&& \forall \dot{v} \forall \dot{w} \in \Sigma^+ :\ [u\sigma = \dot{v}\dot{w}t\ \rightarrow\ \dot{w}t \notin p(E)]
\end{aligned}$$

We prove the implication for each component in the consequent.

- $\sigma \in \Sigma$. It is implied by $|w\sigma| = n + 1$ and $|w| = n$.

- $u \in p(E)$. There is the same component in the antecedent.
- $t \in p(E)$. There is the same component in the antecedent.
- $\exists \dot{s} : \ u\sigma = \dot{s}t$. We know that t is a suffix of $w\sigma$ and that u is a suffix of w. Also, we know that u is the bigger suffix of w that is a proper prefix of a word of E, by $u \in p(E)$ and $\forall x \forall y \in \Sigma^+ : \ w = xyu \rightarrow yu \notin p(E)$. Last, t is the bigger suffix of $w\sigma$ that is a proper prefix of a word of E. Hence t is a suffix of $u\sigma$.
- $\forall \dot{v} \forall \dot{w} \in \Sigma^+ \ [u\sigma = \dot{v}\dot{w}t \ \rightarrow \ \dot{w}t \notin p(E)]$. By the above item, we know that t is a suffix of $u\sigma$. Also, we know that t is the biggest suffix of $w\sigma = xu\sigma$ that is a proper prefix of a word of E. Hence, t is the bigger suffix for the substring $u\sigma$ of $w\sigma$ which is a proper prefix of a word of E.

□

Lemma 23. $w\sigma \in X^{\sigma}_{n+1} \wedge w \in X^{\epsilon}_{n} \rightarrow \sigma \in C^{\sigma}_{\epsilon}$.

Proof. We show that

$$\begin{aligned}
&\text{from } w\sigma \in X^{\sigma}_{n+1} \quad \equiv && w\sigma \in W_{n+1},\ \exists x : \ w\sigma = x\sigma,\ \sigma \in p(E),\\
&&& \forall x \forall y \in \Sigma^+ \ [w\sigma = xy\sigma \rightarrow y\sigma \notin p(E)]\\
&\text{and } w \in X^{\epsilon}_{n} \quad \equiv && w \in W_n,\\
&&& \forall x \forall t \in \Sigma^+ \ [w = xt \ \rightarrow \ t \notin p(E)]\\
&\text{we can conclude } \Longrightarrow \sigma \in C^{\sigma}_{\epsilon} \quad \equiv && \sigma \in \Sigma,\ \sigma \in p(E)
\end{aligned}$$

We prove the implication for each component in the consequent.

- $\sigma \in \Sigma$. This is similar to the proof in the lemma 22.
- $\sigma \in p(E)$. This term is also in the antecedent.

□

Lemma 24. $w\sigma \in X^{\epsilon}_{n+1} \wedge w \in X^{t}_{n} \rightarrow \sigma \in C^{\epsilon}_{t}$.

Proof. We show that

$$\begin{aligned}
&\text{from } w\sigma \in X^{\epsilon}_{n+1} \quad \equiv && w\sigma \in W_{n+1},\\
&&& \forall x \forall u \in \Sigma^+ \ [w\sigma = xu \ \rightarrow \ u \notin p(E)]\\
&\text{and } w \in X^{t}_{n} \quad \equiv && w \in W_n,\ \exists x : \ w = xt,\ t \in p(E),\\
&&& \forall x \forall y \in \Sigma^+ \ [w = xyt \rightarrow yt \notin p(E)]\\
&\text{we can conclude } \Longrightarrow \sigma \in C^{\epsilon}_{t} \quad \equiv && \sigma \in \Sigma,\ t \in p(E),\\
&&& \forall \dot{x} \forall \dot{y} \ [t\sigma = \dot{x}\dot{y}\sigma \ \rightarrow \ \dot{y}\sigma \notin p(E)]
\end{aligned}$$

We prove the implication for each component of the consequent.

- $\sigma \in \Sigma$. It is similar to the proof in the lemma 22.

- $t \in p(E)$. This term is also in the antecedent
- $\forall \dot{x} \forall \dot{y}\ [t\sigma = \dot{x}\dot{y}\sigma\ \rightarrow\ \dot{y}\sigma \notin p(E)]$. We know that every suffix of $w\sigma$ does not is a proper prefix of a word of E, $\forall x \forall u \in \Sigma^+\ [w\sigma = xu\ \rightarrow\ u \notin p(E)]$. Also, we know that $\exists x:\ w = xt$, which imply $w\sigma = xt\sigma$. Hence, every suffix of $t\sigma$ does not is a proper prefix of any word of E.

□

Lemma 25. $w\sigma \in X^{\epsilon}_{n+1} \wedge w \in X^{\epsilon}_{n} \rightarrow \sigma \in C^{\epsilon}_{\epsilon}$.

Proof. We show that

$$\begin{array}{lll} \text{from } w\sigma \in X^{\epsilon}_{n+1} & \equiv & w\sigma \in W_{n+1}, \\ & & \forall x \forall t \in \Sigma^+\ [w\sigma = xt\ \rightarrow\ t \notin p(E)] \\ \text{and } w \in X^{\epsilon}_{n} & \equiv & w \in W_n,\ \forall x \forall t \in \Sigma^+\ [w = xt\ \rightarrow\ t \notin p(E)] \\ \text{we can conclude } \Longrightarrow \sigma \in C^{\epsilon}_{\epsilon} & \equiv & \sigma \in \Sigma,\ \sigma \notin p(E) \end{array}$$

We prove the implication for each term of the consequent.

- $\sigma \in \Sigma$. This is similar to the proof in the lemma 22.
- $\sigma \notin p(E)$. This is a simple case of $\forall x \forall t \in \Sigma^+\ [w\sigma = xt\ \rightarrow\ t \notin p(E)]$ with $t = \sigma$.

□

The following two lemmas show the inductive relation that exists between two consecutive valid-prefix sets.

Lemma 26. *Let $w\sigma$ be a word in X^{t}_{n+1}. Then w belongs to the concatenation of X^{u}_{n} and C^{t}_{u} or belongs to the concatenation of X^{ϵ}_{n} and C^{t}_{ϵ}. That is,*

$$w\sigma \in X^{t}_{n+1}\ \rightarrow\ \exists\, u\colon w\sigma \in X^{u}_{n} \cdot C^{t}_{u} \vee w\sigma \in X^{\epsilon}_{n} \cdot C^{t}_{\epsilon}$$

Proof. By definition, we know that $w\sigma \in X^{t}_{n+1}$ implying that $w \in W_n$. By Lemma 10 we know that $w \in X^{u}_{n} \vee w \in X^{\epsilon}_{n}$. Hence, we can reduce the antecedent to $w\sigma \in X^{t}_{n+1} \wedge w \in X^{u}_{n} \rightarrow \sigma \in C^{t}_{u}$ and $w\sigma \in X^{t}_{n+1} \wedge w \in X^{\epsilon}_{n} \rightarrow \sigma \in C^{t}_{\epsilon}$. We analyze both options.

- $w\sigma \in X^{t}_{n+1} \wedge w \in X^{u}_{n} \rightarrow \sigma \in C^{t}_{u}$. It is proved in the lemma 22
- $w\sigma \in X^{t}_{n+1} \wedge w \in X^{\epsilon}_{n} \rightarrow \sigma \in C^{t}_{\epsilon}$. It is proved in the lemma 23

□

Lemma 27. *Let $w\sigma$ be a word in X^{ϵ}_{n+1}. Then $w\sigma$ belongs to the concatenation of X^{u}_{n} and C^{ϵ}_{u} or belongs to the concatenation of X^{ϵ}_{n} and C^{ϵ}_{ϵ}. That is,*

$$w\sigma \in X^{\epsilon}_{n+1}\ \rightarrow\ \exists\, u\colon w\sigma \in X^{u}_{n} \cdot C^{\epsilon}_{u} \vee w\sigma \in X^{\epsilon}_{n} \cdot C^{\epsilon}_{\epsilon}$$

Proof. By applying the same ideas as in the previous lemma, we study two options.

- $w\sigma \in X^{\epsilon}_{n+1} \wedge w \in X^{t}_{n} \rightarrow \sigma \in C^{\epsilon}_{t}$. It is proved in the lemma 24
- $w\sigma \in X^{\epsilon}_{n+1} \wedge w \in X^{\epsilon}_{n} \rightarrow \sigma \in C^{\epsilon}_{\epsilon}$. It is proved in the lemma 25

□

Lemma 28.

$$X^{t}_{n+1} = X^{\epsilon}_{n} \cdot C^{t}_{\epsilon} \cup_{u} X^{u}_{n} \cdot C^{t}_{u}$$
$$X^{\epsilon}_{n+1} = X^{\epsilon}_{n} \cdot C^{\epsilon}_{\epsilon} \cup_{u} X^{u}_{n} \cdot C^{\epsilon}_{u}$$

Proof. It is necessary to prove the double implication $\forall w, \sigma \in \Sigma^{*}: \ w\sigma \in X^{t}_{n+1} \leftrightarrow w \in \cup_{u} X^{u}_{n} \cdot C^{t}_{u} \ \vee \ \sigma \in X^{\epsilon}_{n} \cdot C^{t}_{\epsilon}$. This is proved by the lemmas 18, 19 and 26.

In the second case, we have to prove the double implication $\forall w, \sigma \in \Sigma^{*}: \ w\sigma \in X^{\epsilon}_{n+1} \leftrightarrow w \in \cup_{u} X^{u}_{n} \cdot C^{\epsilon}_{u} \ \vee \ \sigma \in X^{\epsilon}_{n} \cdot C^{\epsilon}_{\epsilon}$. This is proved by the lemmas 20, 21 and 27. □

Lemma 29.

$$X^{u}_{n} \cdot C^{t}_{u} \cap X^{r}_{n} \cdot C^{t}_{r} = \varnothing$$
$$X^{u}_{n} \cdot C^{\sigma}_{u} \cap X^{\epsilon}_{n} \cdot C^{\sigma}_{\epsilon} = \varnothing$$
$$X^{u}_{n} \cdot C^{\epsilon}_{u} \cap X^{r}_{n} \cdot C^{\epsilon}_{r} = \varnothing$$
$$X^{u}_{n} \cdot C^{\epsilon}_{u} \cap X^{\epsilon}_{n} \cdot C^{\epsilon}_{\epsilon} = \varnothing$$

Proof. We know that $X^{t}_{n} \cap X^{u}_{n} = \varnothing$, with $t \neq u$, and $X^{t}_{n} \cap X^{\epsilon}_{n} = \varnothing$. Hence, the concatenation of the proposed sets are empty because the prefixes of the words are always different. □

The inductive relation between two consecutive valid-prefix sets shown in Lemmas 26 and 27 and the partition shown in Lemmas 28 and 29 drive us to the central result of this paper.

Theorem 30. *The cardinal of a valid-prefix set X_{n+1} is equal to the sum of the cardinals of the valid-prefix sets X_n that compound it.*

$$|X^{t}_{n+1}| = \sum_{u} |X^{u}_{n} \cdot C^{t}_{u}| + |X^{\epsilon}_{n} \cdot C^{t}_{\epsilon}| \tag{15}$$
$$|X^{\epsilon}_{n+1}| = \sum_{u} |X^{u}_{n} \cdot C^{\epsilon}_{u}| + |X^{\epsilon}_{n} \cdot C^{\epsilon}_{\epsilon}| \tag{16}$$

Proof. The union is proved by the lemma 28, and we have proved in lemma 29 that the intersection of the sets is empty. Hence, the cardinal is the sum of the cardinals of the sets. □

Example 31. *Let $E = \{00, 010\}$ the set of avoiding factors in Example 5, valid-prefix sets in Example 9 and linking sets in Example 17. Then we obtain the recurrence relation system*

$$|X^{\epsilon}_{n+1}| = |X^{\epsilon}_{n}| \times |C^{\epsilon}_{\epsilon}| + |X^{01}_{n}| \times |C^{\epsilon}_{01}| = |X^{\epsilon}_{n}| + |X^{01}_{n}|$$
$$|X^{0}_{n+1}| = |X^{\epsilon}_{n}| \times |C^{0}_{\epsilon}| = |X^{\epsilon}_{n}|$$
$$|X^{01}_{n+1}| = |X^{0}_{n}| \times |C^{01}_{0}| = |X^{0}_{n}|$$

with initial terms $|X^{\epsilon}_{2}| = 1$, $|X^{0}_{2}| = 1$ and $|X^{01}_{2}| = 1$.

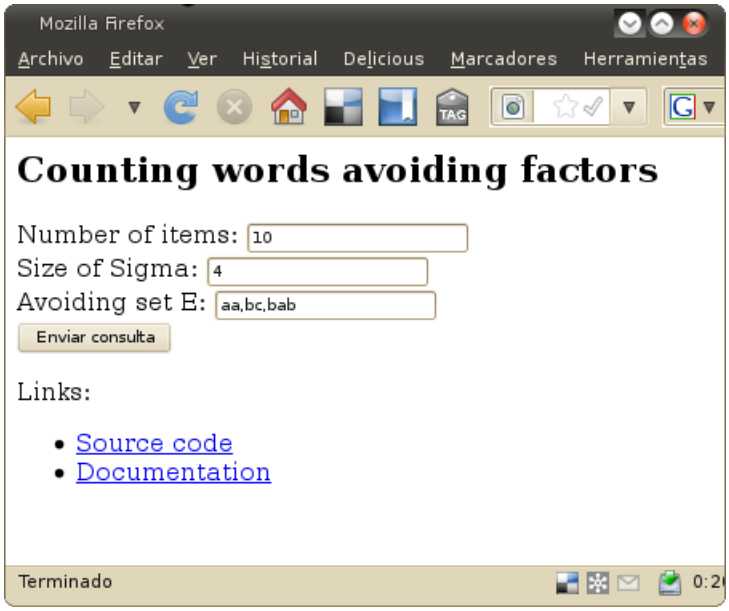

Figure 2: Screenshot of the application configuration for Example 33

Corollary 32. *The number of words avoiding a set of factors $a_E(n)$ can be calculated by counting the cardinal of valid-prefix sets and the cardinal of the linking sets.*

Proof. By the lemma 12 and the theorem 30. □

Example 33. *Let $E = \{00, 010\}$ be the set of avoiding factors in Example 5, valid-prefix sets in Example 9, linking sets in Example 17 and recurrence relation system in Example 31. Then we have*

$$a_E(n+1) = |X^{\epsilon}_{n+1}| + |X^{0}_{n+1}| + |X^{01}_{n+1}|$$

with initial terms $a_E(0) = 1$, $a_E(1) = 2$, $a_E(2) = 3$.

The sequence generated in this sample is $a_E = 1, 2, 3, 4, 6, 9, 13, 19, 28, 41, \ldots$

5 Implementation

An on-line implementation of the solution presented in this paper is available at one of the author's web site[2]. This on-line tool allows students to experiment with our solution, and to verify their own solution. The possibility to interact with the algorithmical solution is very likely to help students to develop their understanding of combinatorics and its methods. Additionally, the source code is available from the author's web site under GNU GPL license, which makes it possible to integrate our solution with other applications.

[2] `http://iaia.lcc.uma.es/~baena/papers/factoravoidance`

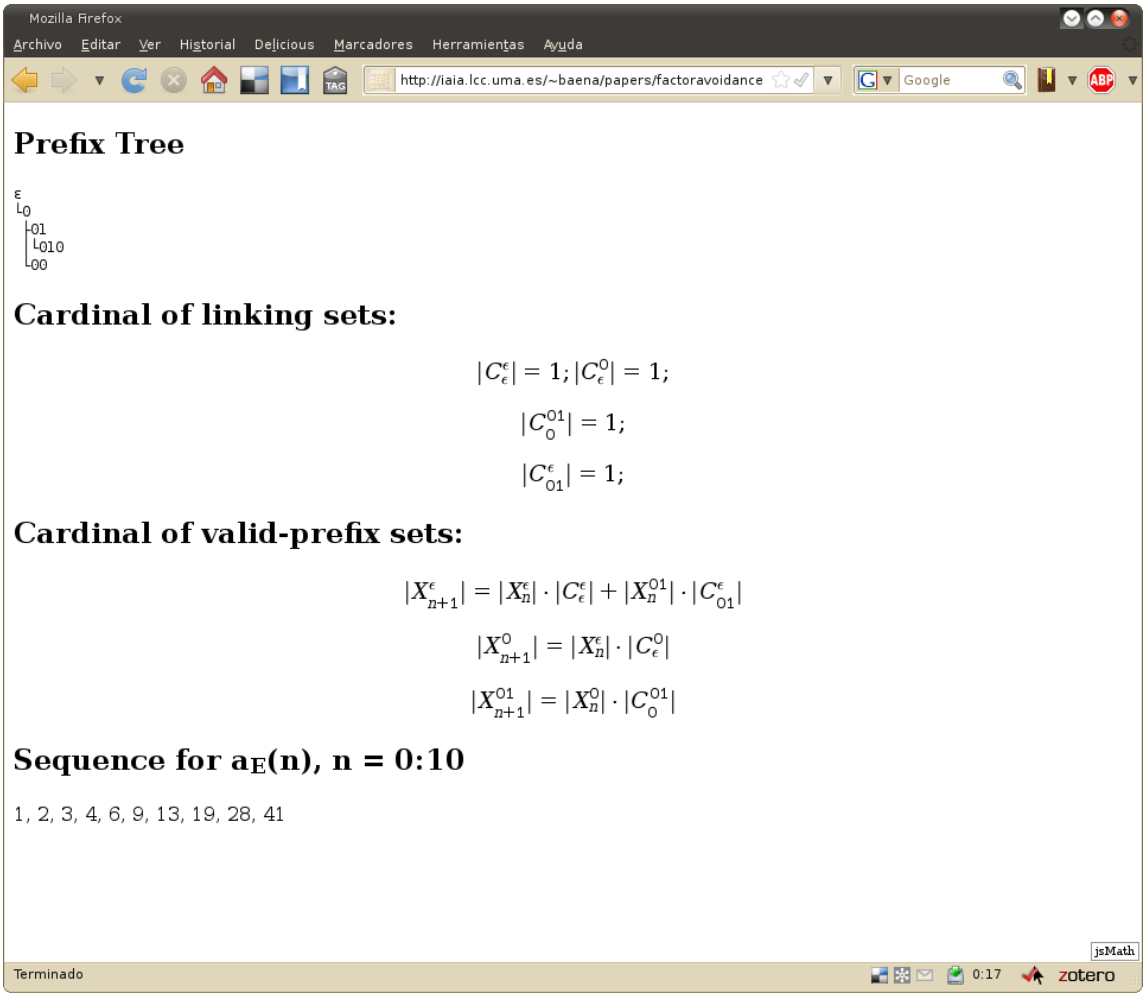

Figure 3: Screenshot of the solution proposed by the on-line application for Example 33

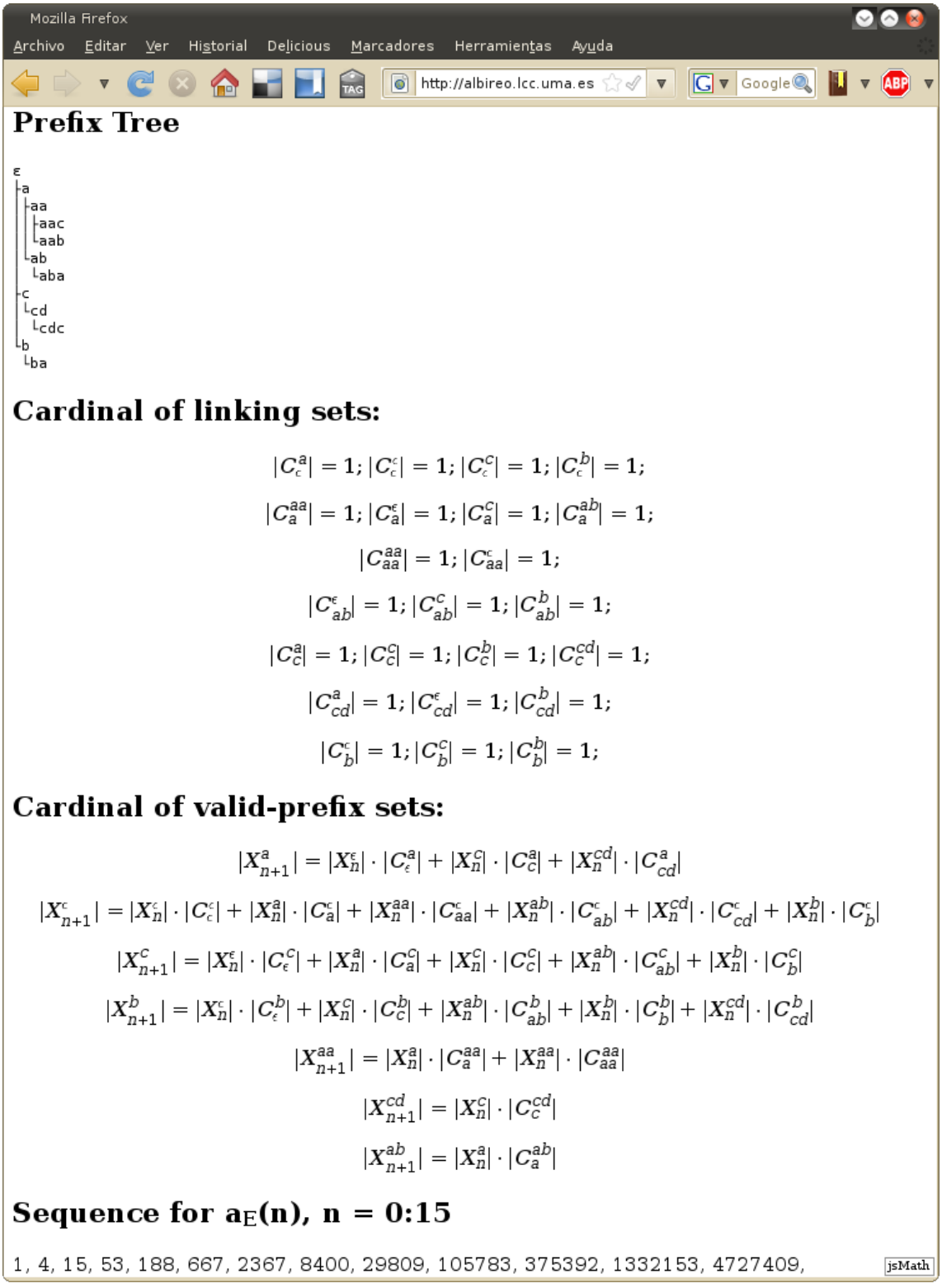

Figure 4: Screenshot of the solution for the problem with the avoiding set $E = \{aac, aab, aba, ba, cdc\}$ and an alphabet with four symbols

6 Conclusions

In this paper we provide a detailed demonstration of the solution of a given combinatorics problem, the *factors avoiding problem*. This demonstration is developed at the level of an undergraduate student. A website has been published with the aim of letting students experiment with an implementation of our solution, enabling them to verify their own work. The source code has been published under the GNU GPL license, letting developers integrate our solution with their own tools.

References

[1] M. H. Albert. On the length of the longest subsequence avoiding an arbitrary pattern in a random permutation. *Random Struct. Algorithms*, 31(2):227–238, 2007.

[2] E. BABSON and E. STEINGRIMSSON. Generalized permutation patterns and a classification of the mahonian statistics, 2000.

[3] Jason P. Bell and Teow Lim Goh. Exponential lower bounds for the number of words of uniform length avoiding a pattern. *Inf. Comput.*, 205(9):1295–1306, 2007.

[4] P.J. Cameron. *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press, 1995.

[5] Timothy Chow and Julian West. Forbidden subsequences and chebyshev polynomials. *Discrete Mathematics*, 204(1-3):119–128, June 1999.

[6] Leonidas J. Guibas and Andrew M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *J. Comb. Theory, Ser. A*, 30(2):183–208, 1981.

[7] S. Kavousian. The Development of Combinatorial Thinking in Undergraduate Students. In *North American Chapter of the International Group for the Psychology of Mathematics Education*, 2005.

[8] M. Lothaire, G.-C. Rota, B. Doran, M. Ismail, T. Y. Lam, E. Wutwak, P. Flajolet, and E. Lutwak. *Applied Combinatorics on Words (Encyclopedia of Mathematics and its Applications)*. Cambridge University Press, New York, NY, USA, 2005.

[9] Marianne Månsson. Pattern avoidance and overlap in strings. *Comb. Probab. Comput.*, 11(4):393–402, 2002.

[10] N. Sloane and T. On. Line encyclopedia of integer sequences.

[11] Zvezdelina E. Stankova. Forbidden subsequences. *Discrete Math.*, 132(1-3):291–316, 1994.

[12] Bridget Eileen Tenner. Pattern avoidance and the bruhat order. *J. Comb. Theory Ser. A*, 114(5):888–905, 2007.

[13] Axel Thue. Über unendliche Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter, I. Mat. Nat. Kl.*, 1906(7):1–22, 1906.

[14] Axel Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter, I. Mat. Nat. Kl.*, 1912(1):1–67, 1912.

[15] Julian West. Generating trees and forbidden subsequences. *Discrete Math.*, 157(1-3):363–374, 1996.

[16] Doron Zeilberger. Enumeration schemes and, more importantly, their automatic generation. *Ann. Comb.*, 2(2):185–195, 1998.

2000 *Mathematics Subject Classification*: 68R15.
Keywords: combinatorial problems, combinatorics on words, factor avoidance.
