

**Biswas, Sudip; Ganguly, Arnab; Shah, Rahul; Thankachan, Sharma V.**

**Ranked document retrieval with forbidden pattern.** (English) [Zbl 1432.68120](#)

Cicalese, Ferdinando (ed.) et al., Combinatorial pattern matching. 26th annual symposium, CPM 2015, Ischia Island, Italy, June 29 – July 1, 2015. Proceedings. Cham: Springer. Lect. Notes Comput. Sci. 9133, 77-88 (2015).

Summary: Let  $\mathcal{D} = \{T_1, T_2, \dots, T_D\}$  be a collection of  $D$  string documents of  $n$  characters in total. The forbidden pattern document listing problem asks to report those documents  $\mathcal{D}' \subseteq \mathcal{D}$  which contain the pattern  $P$ , but not the pattern  $Q$ . The **top- $k$**  forbidden pattern query  $(P, Q, k)$  asks to report those  $k$  documents in  $\mathcal{D}'$  that are most relevant to  $P$ . For typical relevance functions (like document importance, term-frequency, term-proximity), we present a linear space index with worst case query time of  $O(|P| + |Q| + \sqrt{nk})$  for the **top- $k$**  problem. As a corollary of this result, we obtain a linear space and  $O(|P| + |Q| + \sqrt{nt})$  query time solution for the document listing problem, where  $t$  is the number of documents reported. We conjecture that any significant improvement over the results in this paper is highly unlikely.

For the entire collection see [\[Zbl 1314.68012\]](#).

**MSC:**

[68P20](#) Information storage and retrieval of data

[68W32](#) Algorithms on strings

Cited in **2** Documents

**Full Text:** [DOI](#)