As-Soon-As-Possible Top-k Query Processing in P2P Systems

William Kokou Dézoeën, Philippe Lamarre, Reza Akbarinia, and Patrick Valduriez

1 INRIA Rennes, France
William.Dedzoe@inria.fr
2 INSA de Lyon, France
Philippe.Lamarre@insa-lyon.fr
3 INRIA and LIRMM, Montpellier, France
{Reza.Akbarinia,Patrick.Valduriez}@inria.fr

Abstract. Top-k query processing techniques provide two main advantages for unstructured peer-to-peer (P2P) systems. First they avoid overwhelming users with too many results. Second they reduce significantly network resources consumption. However, existing approaches suffer from long waiting times. This is because top-k results are returned only when all queried peers have finished processing the query. As a result, query response time is dominated by the slowest queried peer. In this paper, we address this users’ waiting time problem. For this, we revisit top-k query processing in P2P systems by introducing two novel notions in addition to response time: the stabilization time and the cumulative quality gap. Using these notions, we formally define the as-soon-as-possible (ASAP) top-k processing problem. Then, we propose a family of algorithms called ASAP to deal with this problem. We validate our solution through implementation and extensive experimentation. The results show that ASAP significantly outperforms baseline algorithms by returning final top-k result to users in much better times.

1 Introduction

Unstructured Peer-to-Peer (P2P) systems have gained great popularity in recent years and have been used by millions of users for sharing resources and content over the Internet. In these systems, there is neither a centralized directory nor any control over the network topology or resource placement. Because of few topological constraints, they require little maintenance in highly dynamic environments. However, executing queries over unstructured P2P systems typically by flooding may incur high network traffic and produce lots of query results.

To reduce network traffic and avoid overwhelming users with high numbers of query results, complex query processing techniques based on top-k answers have been proposed e.g. in. With a top-k query, the user specifies a number k of the most relevant answers to be returned by the system. The quality (i.e. score of relevance) of the answers to the query is determined by user-specified scoring...
functions \cite{9,18}. Despite the fact that these top-\(k\) query processing solutions e.g. \cite{2} reduce network traffic, they may significantly delay the answers to users. This is because top-\(k\) results are returned to the user only when all queried peers have finished processing the query. Thus, query response time is dominated by the slowest queried peer, which makes users suffer from long waiting times. Therefore, these solutions are not suitable for emerging applications such as P2P data sharing for online communities, which may have high numbers of autonomous data sources with various access performance. Most of the previous work on top-\(k\) processing has focused on efficiently computing the exact or approximate result sets and reducing network traffic \cite{6,17,34,32,2}.

A naive solution to reduce users’ waiting time is to have each peer return its top-\(k\) results directly to the query originator as soon as it has done executing the query. However, this significantly increases network traffic and may cause a bottleneck at the query originator when returning high numbers of results. In this paper, we aim at reducing users’ waiting time by returning high quality intermediate results, while avoiding high network traffic. The intermediate results are the results of peers which have already processed locally their query. Providing intermediate results to users is quite challenging because a naive solution may saturate users with results of low quality, and incur significant network traffic which in turn may increase query response time.

In this paper, our objective is to return high quality results to users as soon as possible. For this, we revisit top-\(k\) query processing in P2P systems by introducing two notions to complement response time: stabilization time and cumulative quality gap. The stabilization time is the time needed to obtain the final top-\(k\) result set, which may be much lower than the response time (when it is sure that there is no other top-\(k\) result). The quality gap of the top-\(k\) intermediate result set is the quality that remains to be the final top-\(k\) result set. The cumulative quality gap is the sum of the quality gaps of all top-\(k\) intermediate result sets during query execution. Using these notions, we formally define the as-soon-as-possible (ASAP) top-\(k\) processing problem. Then, we propose a family of algorithms called ASAP to deal with this problem.

This paper is an extended version of \cite{12} with the following added value. First, in Section 6 we propose a solution to deal with node failures (or departures) which may decrease the quality and accuracy of top-\(k\) results. In Section 7 we propose two techniques to compute ”probabilistic guarantees” for the users showing for example the probability that current intermediate top-\(k\) results are the true top-\(k\) results (i.e. confidence of current top-\(k\) result). Section 8.2 shows experimentally the effectiveness of our solution for computing “probabilistic guarantees”. Finally, we study experimentally the impact of data distribution on our algorithms (Section 8.2).

2 System Model

In this section, we first present a general model of unstructured P2P systems which is needed for describing our solution. Then, we provide a model and definitions for top-\(k\) queries.