The DESQ Framework for Declarative and Scalable Frequent Sequence Mining

Presentation of work originally published in IEEE 16th Intl. Conf. on Data Mining, IEEE 35th Intl. Conf. on Data Engineering, and 2019 ACM Trans. on Database Syst.

Kaustubh Beedkar[1], Rainer Gemulla[2], Alexander Renz-Wieland[3]

Abstract: DESQ is a general-purpose framework for declarative and scalable frequent sequence mining. Applications express their specific sequence mining tasks using a simple yet powerful powerful pattern expression language, and DESQ's computation engine automatically executes the mining task in an efficient and scalable way. In this paper, we give a brief overview of DESQ and its components.

Keywords: Data mining; Frequent sequence mining; Subsequence constraints; Pattern expressions; Distributed sequence mining

Frequent sequence mining (FSM, [Do09]) is a fundamental task in data mining: Given a sequence database, the task is to find interesting sequential patterns that appear frequently in the data. In customer behavior analysis, for example, frequent sequences may correspond to purchase patterns of customers, or to popular navigations paths across a website, and can serve as an input for applications such as recommender systems. The task arises in a wide range of applications, including natural language processing, information extraction, web usage mining, market-basket analysis, and computational biology.

DESQ is a general-purpose framework for FSM that aims to support such a wide range of applications. DESQ features (1) a powerful pattern expression language that allows applications to describe declaratively which patterns are considered interesting, (2) a suite of efficient and scalable mining algorithms that support both sequential and distributed execution, and (3) an easy-to-use API based on Apache Spark. The DESQ framework allows applications to express a wide range of sequence mining problems—including and beyond those considered in prior literature—in a unified way. This unified treatment improves the usability of pattern mining in practice: Data scientists only need to familiarize themselves with one framework and, perhaps more importantly, do not need to develop customized mining algorithms for a particular application. Likewise, a unified treatment allows researchers to study jointly many variants of FSM, instead of each one individually.

Consider for example the task of mining frequent relational phrases between entities from large text corpora; e.g., the phrase *make a deal with* may be frequent between persons and/or organizations. Such patterns are indicative of relations between entities and arise in natural language processing and information extraction applications. Existing

[1] Technische Universität Berlin, kaustubh.beedkar@tu-berlin.de
[2] Universität Mannheim, rgemulla@uni-mannheim.de
[3] Technische Universität Berlin, alexander.renz-wieland@tu-berlin.de

doi:10.18420/inf2019_28

Tab. 1: Example pattern expressions for some FSM tasks and frequencies in New York Times data (first two blocks) and Amazon review data (last block).

Pattern expression	FSM task	Example patterns (frequency)
(.){1,4}	*n-grams of up to four words*	green tea (337), editor in chief (3275)
(.)[.{0,2}(.)]{1,3}	*Skip n-grams of 2–4 words with gap at most 2*	flight from to (758), son of and of (15896)
ENTITY (VERB$^+$ DET? NOUN$^+$? PREP?) ENTITY	*Relational phrases*	is being advised by (15), has coached (10)
(ENTITY$^\uparrow$ VERB$^+$ NOUN$^+$? PREP? ENTITY$^\uparrow$)	*Typed relational phrases*	ORG headed by ENTITY (275), PER born in LOC (481)
(Book)[.{0,2}(Book)]{1,4}	*Sequences of books*	'A Storm of Swords' 'A Feast for Crows'(153)
DigitalCamera[.{0,3}(.$^\uparrow$)]{1,4}	*Products or types of products purchased after a digital camera*	'Lenses' 'Tripods'(158), 'Batteries' 'SD&SDHC Cards'(149)

FSM algorithms cannot solve such a task since they cannot be tailored to consider only relational phrases (thereby producing many uninteresting—i.e., non-relational—patterns) or to consider context information (thereby producing patterns that do not connect entities). In contrast, this mining task can be expressed in DESQ's pattern expression language as *ENTITY* (*VERB*$^+$ *DET*? *NOUN*$^+$? *PREP*?) *ENTITY*.

DESQ's pattern expression language is based on regular expressions and additionally includes features such as item hierarchies and capture groups. In the above example, item hierarchies allow applications to relate items to each other (e.g., *make* is a *VERB*), and capture groups allow to express what is considered part of the pattern (the relational phrase) and what context (between entities). Table 1 lists some additional examples, in which pattern expressions are used to either concisely describe traditional FSM tasks or to define customized sequence mining tasks.

DESQ includes a number of general-purpose mining algorithms for the wide range of mining tasks that can be expressed using pattern expressions. In particular, DESQ provides efficient algorithms that can operate on a single machine as well as scalable, distributed algorithms. The pattern expression language, the underlying computational framework, and efficient mining algorithms are described in [BG16; BGM19; RBG19].

DESQ is available as an open source library.[4] The library provides a Java API for a single machine setup, and a Scala API for a distributed setup on top of Apache Spark. The API allows applications to perform pattern mining directly on datasets in their native formats.

Literatur

[BG16] Beedkar, K.; Gemulla, R.: DESQ: Frequent Sequence Mining with Subsequence Constraints. In: ICDM. S. 793–798, 2016.

[BGM19] Beedkar, K.; Gemulla, R.; Martens, W.: A Unified Framework for Frequent Sequence Mining with Subsequence Constraints. ACM Trans. Database Syst. 44/3, 11:1–11:42, 2019.

[Do09] Dong, G.: Sequence Data Mining. Springer-Verlag, Berlin, Heidelberg, 2009.

[RBG19] Renz-Wieland, A.; Bertsch, M.; Gemulla, R.: Scalable Frequent Sequence Mining With Flexible Subsequence Constraints. In: ICDE. S. 1490–1501, 2019.

[4] `https://www.uni-mannheim.de/dws/research/resources/desq`