Li *et al. J Big Data* (2021) 8:30
https://doi.org/10.1186/s40537-021-00416-y

Journal of Big Data

RESEARCH Open Access

A sample decreasing threshold greedy-based algorithm for big data summarisation

Teng Li, Hyo-Sang Shin* and Antonios Tsourdos

*Correspondence:
h.shin@cranfield.ac.uk
School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield MK43 0AL, UK

Abstract

As the scale of datasets used for big data applications expands rapidly, there have been increased efforts to develop faster algorithms. This paper addresses big data summarisation problems using the submodular maximisation approach and proposes an efficient algorithm for maximising general non-negative submodular objective functions subject to k-extendible system constraints. Leveraging a random sampling process and a decreasing threshold strategy, this work proposes an algorithm, named Sample Decreasing Threshold Greedy (SDTG). The proposed algorithm obtains an expected approximation guarantee of $\frac{1}{1+k} - \epsilon$ for maximising monotone submodular functions and of $\frac{k}{(1+k)^2} - \epsilon$ in non-monotone cases with expected computational complexity of $O\left(\frac{n}{(1+k)\epsilon} \ln \frac{r}{\epsilon}\right)$. Here, r is the largest size of feasible solutions, and $\epsilon \in \left(0, \frac{1}{1+k}\right)$ is an adjustable designing parameter for the trade-off between the approximation ratio and the computational complexity. The performance of the proposed algorithm is validated and compared with that of benchmark algorithms through experiments with a movie recommendation system based on a real database.

Keywords: Big data summarisation, Submodular maximisation, k-extendible system constraints, Personalised recommendation

Introduction

The research of big data has received extensive attention due to its great significance [1]. Data summarisation, which involves extracting representative information with certain constraints from a large-scale dataset, is one of the compelling directions of big data processing [2]. Typical applications of big data summarisation include personalised recommendation systems [3–6], exemplar-based clustering [7–9], and summarisation of text [10, 11], images [12–14], corpus [8, 15], and videos [16, 17], just to name a few.

The unprecedented growth of modern datasets requires efficient and effective techniques to process a mass of data. Computational complexity is one of the grand challenges of big data operations [1]. Fortunately, the quality of data summarisation outcome can be often measured by submodular set functions [11, 12, 14], where the marginal gain value of an element decreases as more elements have already been

selected, namely diminishing returns [18]. It is well known that the greedy-related algorithms are efficient and can provide an approximation guarantee for maximising submodular functions [19]. Hence, the big data summarisation problem can be handled as maximising a submodular function based on a large-scale dataset, meanwhile satisfying a certain constraint or a combination of several constraints [2].

This paper addresses big data summarisation problems using the submodular maximisation approach, especially subject to k-extendible system constraints. Note that the k-extendible system constraint is a general type of constraint that has been widely studied. The concept of k-extendible systems was first introduced by Mestre in 2006 [20]. The intersection of k matroids based on the same ground set is always k-extendible [20]. Many types of constraints handled in submodular maximisation problems fall into the k-extendible system constraint, such as the cardinality constraint, partition matroid constraint, and k-matroid constraint.

The issue is that finding the optimal solution of submodular maximisation is NP-hard, and the sizes of datasets tend to increase. NP-hard problems are known to significantly suffer from "curse of dimensionality", which implies that the complexity of the problem explodes as the problem size increases. Therefore, the trend of increasing sizes of datasets combined with the NP-hardness of the problem urges the development of more computationally efficient optimisation algorithms. The Sample Greedy algorithm (Sample, for short) proposed in [21] is one of the state-of-the-art algorithms for constrained submodular maximisation problems. Specifically, Sample [21] was the fastest algorithm (before this work) for maximising non-monotone submodular functions subject to a k-extendible system constraint.

Inspired by the sampling strategy from [21] and a decreasing threshold idea from [22], this work proposes an algorithm that is even faster than Sample [21]. The proposed algorithm, which is named as Sample Decreasing Threshold Greedy (SDTG), provides an expected approximation guarantee of $p - \epsilon$ for maximising monotone submodular functions and of $p(1-p) - \epsilon$ for non-monotone cases with expected time complexity of only $O(\frac{pn}{\epsilon} \ln \frac{r}{\epsilon})$, where $p \in (0, \frac{1}{1+k}]$ is the sampling probability and $\epsilon \in (0, p)$ is the threshold decreasing parameter. If the sampling probability p is set as $\frac{1}{1+k}$, then SDTG provides the best approximation ratios for both monotone and non-monotone submodular functions which are $\frac{1}{1+k} - \epsilon$ and $\frac{k}{(1+k)^2} - \epsilon$, respectively. Here, ϵ acts as a design parameter for the trade-off between the approximation ratio and the computational complexity. The proposed algorithm is validated through experiments with a movie recommendation system based on the MovieLens [23] which is a widely used real movie information database. Experimental results demonstrate that the proposed algorithm outperforms benchmark algorithms in terms of both solution quality and computation efficiency. The main contributions of this work are summarised as follows:

- This work proposes the current *fastest* algorithm, SDTG, for maximising non-monotone submodular functions subject to k-extendible system constraints;
- Precise mathematical proofs are provided for analysing the theoretical guarantees of the proposed algorithm;

- Experiments with a movie recommendation system based on a real database are carried out to reveal the practical performance of SDTG for solving the big data summarisation problem.

The rest part of this work is organised as follows. "Related works" section investigates related articles for constrained submodular maximisation problems. In "Preliminaries" section, some basic knowledge related to the proposed algorithm is presented. "Algorithm and analysis" section demonstrates the proposed algorithm and analyses its theoretical performance in detail. The performance and validity of the theoretical results are then testified through experiments with a movie recommendation system in "Experiments" section. "Conclusions" section offers the conclusions of this paper and possible future research directions.

Related works

There have been numerous works recently carried out to develop more efficient constrained submodular maximisation algorithms, and many of them endeavour to increase computational efficiency even by sacrificing some degree of approximation ratio. These works are classified by the types of constraints, and their developments are summarised in the following.

Cardinality constraint

The Sieve-Streaming proposed by Badanidiyuru et al. [12] is the first single-pass streaming algorithm for maximising monotone submodular functions, achieving approximation guarantee of $1/2-\epsilon$ with computational complexity of $O(\frac{n}{\epsilon}\log r)$. Here, n is the size of the ground set, r is the size of the largest feasible solution. Norouzi-Fard et al. [9] proposed another single-pass algorithm Salsa that improved the approximation guarantee to a value better than 1/2. They also extended their work to a multi-pass algorithm P-Pass that provided the trade-off between the approximation ratio and the number of passes. The Decreasing Threshold Greedy proposed in [22] obtained an approximation ratio of $1-1/e-\epsilon$ with time complexity of $O(\frac{n}{\epsilon}\log\frac{n}{\epsilon})$ for monotone submodular functions. This is the first streaming algorithm whose computational complexity is independent of r. Later, the sampling-based Stochastic Greedy proposed by Mirzasoleiman et al. [24] achieved an expectantly the same approximation ratio with lower time complexity of $O(n\log\frac{1}{\epsilon})$, compared with the Decreasing Threshold Greedy [22]. The Stochastic Greedy gets orders of magnitudes faster by losing only a bit of approximation ratio compared with other benchmark algorithms. Then Buchbinder et al. [25] extended the Stochastic Greedy to general non-monotone cases and achieved an approximation guarantee of $1/e-\epsilon$ with computational complexity of $O(\frac{n}{\epsilon^2}\log\frac{1}{\epsilon})$. Recently, Breuer et al. [26] proposed an efficient algorithm Fast for the monotone case, using the adaptive sequencing technique. Fast achieves an approximation ratio of $1-1/e-\epsilon$, with $O(n\log\log r)$ queries.

Matroid constraint

The original greedy algorithm (Greedy) [19] provides an approximation ratio of 1/2 with time complexity of $O(nr)$ for monotone submodular maximisation. Nemhauser and Wolsely [27] proved that no algorithm can achieve an approximation ratio better than

$1-1/e$ with polynomial time complexity. The continuous greedy based on the multilinear extension was utilised to achieve an approximation ratio of $1-1/e$ [28]. The measured continuous greedy algorithm developed by Feldman et al. [29] achieved a $(1-1/e)$ -approximation for the monotone case and a $1/e$-approximation for the non-monotone case. This is the first algorithm to provide a constant factor of approximation for maximising non-monotone submodular functions subject to a partition matroid constraint. However, the sophisticated continuous algorithms are inherently too time-consuming to be applied directly in the real world [30]. To remedy this, the idea of decreasing threshold [22] was adapted to reduce the computational complexity [31]. Badanidiyuru and Vondrak [22] proposed a new variant of the continuous greedy algorithm and achieved an approximation ratio of $1-1/e-\epsilon$ with complexity of $O(\frac{nr}{\epsilon^4}\log^2\frac{r}{\epsilon})$ for monotone submodular functions. Then, a close variant of the Decreasing Threshold Greedy described in [25] provided an approximation ratio of $1/2-\epsilon$ with computational complexity of $O(\frac{n}{\epsilon}\log\frac{r}{\epsilon})$ for the monotone case.

k-extendible system constraint

It is known that Greedy [19] achieves a $\frac{1}{1+k}$-approximation for maximising monotone submodular functions subject to a k-extendible system constraint. The Decreasing Threshold Greedy [22] provides a slightly worse approximation guarantee of $\frac{1}{1+k+\epsilon}$ but requires lower computational complexity of $O(\frac{n}{\epsilon^2}\log^2\frac{n}{\epsilon})$ than Greedy [19] does for maximising monotone submodular functions. For the non-monotone case, Gupta et al. [32] proposed an algorithm achieving an approximation ratio of $\frac{k}{(k+1)(3k+3)}$ with time complexity of $O(nrk)$. Then, the approximation ratio was improved to $\frac{k}{(k+1)(2k+1)}$ by an algorithm called Fantom proposed by Mirzasoleiman et al. [5] with the same complexity. After this, Feldman et al. [21] made a significant breakthrough in terms of both approximation ratio and time complexity. The Sample algorithm proposed in [21] achieved an approximation ratio of $\frac{k}{(k+1)^2}$ with complexity of $O(n+nr/k)$. Experiments based on a movie recommendation system in [21] confirmed that Sample outperformed Fantom in terms of computational efficiency.

In summary, gradual improvements have been made for solving the constrained submodular maximisation problems recently. However, the rapid expansion in the scale of modern datasets urges persistent developments for faster algorithms. An immediate research question would be whether or not one can develop an algorithm that can further improve the efficiency of maximising general non-negative submodular functions especially subject to k-extendible system constraints.

Preliminaries

This section presents some necessary definitions and basic concepts related to the proposed algorithm. The definitions and concepts can also be found in our previous works [33–35].

Definition 1 (*Submodularity* [21]) A set function $f: 2^{\mathcal{N}} \to \mathbb{R}$ is *submodular* if, $\forall X, Y \subseteq \mathcal{N}$,

$$f(X)+f(Y) \geq f(X \cap Y)+f(X \cup Y).$$

where $\mathcal{N}$ is named as "ground set" which is a finite set containing all elements. Equivalently, $\forall A \subseteq B \subseteq \mathcal{N}$ and $u \in \mathcal{N} - B$,

$$f(A \cup \{u\}) - f(A) \geq f(B \cup \{u\}) - f(B). \tag{1}$$

Definition 2 (*Marginal gain value* [36] *(mgv)*) For a set function $f : 2^{\mathcal{N}} \to \mathbb{R}$, a set $S \subseteq \mathcal{N}$, and an element $u \in \mathcal{N}$, the *marginal gain value* of f at S with respect to u is defined as

$$\Delta f(u|S) \doteq f(S \cup \{u\}) - f(S),$$

where $\doteq$ means equal by definition. This work denotes the *marginal gain value* as "*mgv*" for tidiness.

The inequality (1) is known as the *diminishing return*, which is a crucial property of submodular functions: the *mgv* of a given element will never increase as more elements have already been selected. One intuitive example for the submodularity is the sensor placement problem: The space coverage increment obtained by adding an extra fire detector to a particular position of a room will never increase as more detectors have already been placed in the room.

Definition 3 (*Monotonicity* [36]) A set function $f : 2^{\mathcal{N}} \to \mathbb{R}$ is *monotone* if, $\forall A \subseteq B \subseteq \mathcal{N}$, $f(A) \leq f(B)$. f is *non-monotone* if it is not monotone.

The submodular objective functions considered in this paper are normalised (i.e. $f(\emptyset) = 0$), non-negative (i.e. $f(S) \geq 0$, $\forall S \subseteq \mathcal{N}$), and can be either monotone or non-monotone.

Definition 4 (*Matroid* [22]) A matroid is a pair $\mathcal{M} = (\mathcal{N}, \mathcal{I})$ where $\mathcal{N}$ is the ground set, and $\mathcal{I} \subseteq 2^{\mathcal{N}}$ is a collection of independent sets, satisfying:

- $\emptyset \in \mathcal{I}$;
- If $A \subseteq B, B \in \mathcal{I}$, then $A \in \mathcal{I}$;
- If $A, B \in \mathcal{I}, |A| < |B|$, then $\exists\, u \in B - A$ such that $A \cup \{u\} \in \mathcal{I}$.

Specifically, matroid constraints include uniform matroid constraints and partition matroid constraints. The uniform matroid constraint is also called *cardinality constraint*, which is a special case of matroid constraints where any subset $S \subseteq \mathcal{N}$ satisfying $|S| \leq r$ is independent, i.e. $S \in \mathcal{I}$. The *partition matroid constraint* means that an independent subset S can contain at most a certain number of elements from each of the disjoint partitions of $\mathcal{N}$.

A typical example for the partition matroid constraint is the security camera system: Each camera of the system can only point to one of its admissible directions at a certain moment. The partition matroid constraint is a special case of k-extendible system constraints where k equals to 1. A formal definition of the k-extendible system constraint is given following an auxiliary concept.

Definition 5 (*Extension* [21]) If an independent set B strictly contains an independent set A, then B is called an extension of A.

Definition 6 (*k-extendible system* [20]) A k-extendible system is an independence system $(\mathcal{N}, \mathcal{I})$ that for every independent set $A \in \mathcal{I}$, an extension B of A, and an element $u \notin A$, $A \cup \{u\} \in \mathcal{I}$, there exists a subset $X \subseteq B - A$ with $|X| \leq k$ such that $(B - X) \cup \{u\} \in \mathcal{I}$.

Intuitively, if an element u is added into an independent set A of a k-extendible system, it requires at most k other elements to be removed from A in order to keep the set independent [21]. For example, a certain user of a movie recommendation system likes three genres of movies: Action, Adventure, and Sci-Fi. Suppose that this user wants at most *one* movie from each of these three genres. Note that a movie can belong to multiple genres. Here are four movies with genre information: mv_1 (Action), mv_2 (Adventure), mv_3 (Sci-Fi), and mv_4 (Action, Adventure, Sci-Fi). According to the requirement from the user, a recommendation list $S = \{mv_1, mv_2, mv_3\}$ is independent, i.e., $S \in \mathcal{I}$; adding mv_4 to S will make it dependent. Movies mv_1, mv_2, and mv_3 must be removed from S to keep it independent if mv_4 is remained in S. Therefore, the constraint in this example is a 3-extendible system constraint.

The following is an important claim that provides the mathematical foundation for Sample [21] to work well in non-monotone submodular maximisation. Readers are referred to [37] for the proof of Claim 1.

Claim 1 (Due to [37]) *Let $h : 2^{\mathcal{N}} \to \mathbb{R}_{\geq 0}$ be a submodular function, and let S be a random subset of $\mathcal{N}$. If each element of S appears with a probability at most p (not necessarily independently), then $\mathbb{E}[h(S)] \geq (1 - p)h(\emptyset)$.*

Algorithm and analysis

This section describes SDTG in Algorithm 1 and analyses its theoretical performance in detail. Note that the proposed algorithm is based on submodular optimisation like in our previous studies [33–35]. Hence the analysis shares some essences of logic in our previous works. An equivalent version of Algorithm 1 is introduced as Algorithm 2 to better analyse SDTG.

Algorithm

This work proposes to leverage the sampling strategy [21] and develop a variant of decreasing threshold idea to design a summarisation algorithm. On the one hand, the random sampling at the beginning of SDTG can help the algorithm to avoid getting trapped in local optima. It can also help to accelerate the algorithm because only a small portion of elements from the ground set is considered. On the other hand, the decreasing threshold can further accelerate the algorithm. Note that Greedy [19] needs to reevaluate all the remaining elements to find the best one during each iteration. In contrast, SDTG searches for a relatively good element whose *mgv* is no less

than the current threshold instead of looking for the best one. Therefore, SDTG does not have to reevaluate all remaining elements every time before selecting an extra element.

Some notations from Algorithm 1 are stated in the following: $\mathcal{N}$ is the ground set containing all elements. $\mathcal{I}$ is the collection of all feasible sets (independent); r is the maximum cardinality of feasible sets in $\mathcal{I}$; p is the sampling probability (uniform distribution); ϵ is the threshold decreasing parameter determining the decreasing speed of the threshold; S is the solution set containing the selected elements; R is a set containing the remaining sampled elements; θ is the decreasing threshold.

The structure of Algorithm 1 consists of two phases. The first phase (lines 1–4) is sampling where elements are randomly selected from the ground set $\mathcal{N}$ with probability p to form a sample set R. The probability distribution of sampling is uniform. The second phase (lines 5–22) is selecting where an independent solution set S is selected from R using decreasing threshold greedy. The initial threshold is set as the largest *mgv* given the empty set and denoted as d (line 5). The terminal threshold is set as $\frac{\epsilon}{r}d$ (line 6). The reason for choosing this value as the termination condition will be given later in the proof part.

Algorithm 1 SDTG

Input: $f : 2^{\mathcal{N}} \to \mathbb{R}_{\geq 0}, \mathcal{N}, \mathcal{I}, r, p, \epsilon$.
Output: A set $S \in \mathcal{I}$.

```
1: S ← ∅, R ← ∅
2: for u ∈ N do
3:     R ← R ∪ {u} with probability p                    // Random sampling.
4: end for
5: d ← max_{u∈R} Δf(u|S)
6: for (θ = d; θ ≥ (ε/r)d; θ ← θ(1 − ε)) do
7:     for u ∈ R do
8:         if S ∪ {u} ∉ I then                          // Check independence.
9:             R ← R − {u}
10:        else
11:            if Δf(u|S) ≥ θ then
12:                S ← S ∪ {u}
13:                R ← R − {u}
14:            else
15:                if Δf(u|S) < (ε/r)d then
16:                    R ← R − {u}
17:                end if
18:            end if
19:        end if
20:    end for
21: end for
22: return S
```

More details of the second phase are given in the following. One loop of the inner "**for**" loops is named as one *iteration*. At the beginning of each iteration, SDTG checks independency of $S \cup \{u\}$. If it is not independent, then remove element u from R (lines 8–9). Otherwise, calculate the *mgv* of u and compare it with the current threshold θ. If the *mgv* of u is greater than or equals to θ, then add u to S and remove it from R (lines 11–13). An element u is named as a *qualified element* if the *mgv* of u given S is no less than the current threshold θ. If the *mgv* of an element is already less than $\frac{\epsilon}{r}d$,

it will never become greater or equal to $\frac{\epsilon}{r}d$ in subsequent iterations due to submodularity. Therefore, this element can be removed from R immediately, as stated in lines 15–17. Note that each element in R will be evaluated only for one time under one threshold. If the *mgv* of an element is between $\frac{\epsilon}{r}d$ and θ, this element will remain in R for the next outer loop where the threshold will decrease. The remaining elements in R will be reevaluated and their updated *mgv*s will be compared with a decreased new threshold. The threshold keeps decreasing after all remaining elements in R have been evaluated until reaching the termination condition.

Analysis

To better analyse the theoretical approximation performance of Algorithm 1, this work leverages some analysing techniques that were used in [21]. A few auxiliary variables have been introduced to transform SDTG to an equivalent version, i.e., Algorithm 2.

Algorithm 2 Equivalent SDTG

Input: $f : 2^{\mathcal{N}} \rightarrow \mathbb{R}_{\geq 0}, \mathcal{N}, \mathcal{I}, r, p, \epsilon$.
Output: A set $S \in \mathcal{I}$.

1: $S \leftarrow \emptyset$, $\mathcal{N}_s \leftarrow \emptyset$, $R \leftarrow \mathcal{N}$,
2: $C \leftarrow \emptyset$, $Q \leftarrow OPT$
3: **for** $u \in R$ **do**
4: $\quad \mathcal{N}_s \leftarrow \mathcal{N}_s \cup \{u\}$ with probability p
5: **end for**
6: $d \leftarrow \max_{u \in \mathcal{N}_s} \Delta f(u|S)$
7: **for** $(\theta = d; \theta \geq \frac{\epsilon}{r}d; \theta \leftarrow \theta(1-\epsilon))$ **do**
8: $\quad$ **for** $u \in R$ **do**
9: $\quad\quad$ **if** $S \cup \{u\} \notin \mathcal{I}$ **then**
10: $\quad\quad\quad R \leftarrow R - \{u\}$
11: $\quad\quad$ **else**
12: $\quad\quad\quad$ **if** $\Delta f(u|S) \geq \theta$ **then**
13: $\quad\quad\quad\quad c \leftarrow u$
14: $\quad\quad\quad\quad S_c \leftarrow S$
15: $\quad\quad\quad\quad C \leftarrow C \cup \{c\}$
16: $\quad\quad\quad\quad R \leftarrow R - \{c\}$
17: $\quad\quad\quad\quad$ **if** $u \in \mathcal{N}_s$ **then**
18: $\quad\quad\quad\quad\quad S \leftarrow S \cup \{c\}$
19: $\quad\quad\quad\quad\quad Q \leftarrow Q \cup \{c\}$
20: $\quad\quad\quad\quad\quad$ Let $K_c \subseteq Q - S$ be the smallest set s.t. $Q - K_c \in \mathcal{I}$
21: $\quad\quad\quad\quad$ **else**
22: $\quad\quad\quad\quad\quad$ **if** $c \in Q$ **then**
23: $\quad\quad\quad\quad\quad\quad K_c \leftarrow \{c\}$
24: $\quad\quad\quad\quad\quad$ **else**
25: $\quad\quad\quad\quad\quad\quad K_c \leftarrow \emptyset$
26: $\quad\quad\quad\quad\quad$ **end if**
27: $\quad\quad\quad\quad$ **end if**
28: $\quad\quad\quad\quad Q \leftarrow Q - K_c$
29: $\quad\quad\quad$ **else**
30: $\quad\quad\quad\quad$ **if** $\Delta f(u|S) < \frac{\epsilon}{r}d$ **then**
31: $\quad\quad\quad\quad\quad R \leftarrow R - \{u\}$
32: $\quad\quad\quad\quad$ **end if**
33: $\quad\quad\quad$ **end if**
34: $\quad\quad$ **end if**
35: $\quad$ **end for**
36: **end for**
37: **return** S

In Algorithm 2, variables C, S_c, Q, and K_c are introduced only for the convenience of analysis and have no effect on the final output S. Therefore, Algorithm 2 and Algorithm 1 are equivalent in terms of solution quality. The rules of these variables are as follows.

C is a set that contains all considered elements that have *mgv*s greater or equal to the threshold θ in a certain iteration of Algorithm 2 no matter whether they are added into S or not.

S_c is a set that contains the selected elements at the beginning of the current iteration. At the end of this iteration, $S = S_c \cup \{c\}$ if c is added into S and Q, otherwise S equals to S_c.

Q is a set that bridges the relationship between the solution S and the optimal solution OPT. Q starts at OPT at the beginning of the algorithm and changes over time. Note that, Q is introduced only for analysis and there is no need to know the exact value of Q or OPT. In each iteration, the element added into S is also added into Q. At the same time, a set K_c is removed from Q to keep the independence of Q if an element c is added into Q. Note that, if an element c is already in Q and is considered but not added into S at the current iteration, then this element c should be removed from Q.

K_c is a set that is introduced to keep Q independent and help Q to remove c that is not added to S. According to the property of k-extendible systems, Algorithm 2 is able to remove a set $K_c \subseteq Q - S$ which contains at most k elements from Q if an element is added into the currently independent set Q. In addition, if c is not added to S and $c \in Q$ at the beginning of some iteration, then $K_c = \{c\}$.

The theoretical performance of the proposed algorithm SDTG is summarised in Theorem 1.

Theorem 1 *SDTG achieves an approximation guarantee of at least* $\frac{1}{1+k} - \epsilon$ *for maximising monotone submodular functions subject to k-extendible system constraints and of* $\frac{k}{(1+k)^2} - \epsilon$ *for non-monotone cases with computational complexity of* $O(\frac{n}{(1+k)\epsilon} \ln \frac{r}{\epsilon})$, *where n sis the size of the ground set, r is the largest size of a feasible solution, and* $\epsilon \in (0, \frac{1}{1+k})$ *is the threshold decreasing parameter.*

The computational complexity can be easily proved. Assume that there are in total x number of loops in the outer "**for**" loop of Algorithm 1. Thus,

$$(1-\epsilon)^x = \frac{\epsilon}{r}.$$

Solving the above equation yields

$$x = \frac{\ln \frac{r}{\epsilon}}{\ln \frac{1}{1-\epsilon}} \leq \frac{1}{\epsilon} \ln \frac{r}{\epsilon}.$$

There are expectantly at most $p \cdot n$ function evaluations in each outer loop. Therefore, the time complexity of Algorithm 1 is $O(\frac{pn}{\epsilon} \ln \frac{r}{\epsilon})$. □

The following part of this section analyses the approximation ratios of SDTG in both monotone and non-monotone cases through Algorithm 2.

Lemma 1 $f(S) > \frac{1}{1+\epsilon} f(Q)$.

Proof According to Algorithm 2, at the end of each iteration, the set Q is independent i.e. $Q \in \mathcal{I}$. S is a subset of Q, i.e. $S \subseteq Q$, as every element c that is added to S is also in Q. Therefore, $S \cup \{q\} \in \mathcal{I}\ \forall q \in Q - S$ by the property of independent systems and $|Q - S| \leq r$. At the termination of Algorithm 2, $\Delta f(q|S) < \frac{\epsilon}{r} d\ \forall q \in Q - S$ and $f(S) \geq d$. Thus,

$$\sum_{q \in Q-S} \Delta f(q|S) < \sum_{q \in Q-S} \frac{\epsilon}{r} d \leq \epsilon \cdot \frac{|Q-S|}{r} f(S) \leq \epsilon \cdot f(S).$$

Let $Q - S = \{q_1, q_2, \ldots, q_{|Q-S|}\}$, then

$$\begin{aligned} f(S) &= f(Q) - \sum_{i=1}^{|Q-S|} \Delta f(q_i | S \cup \{q_1, \ldots, q_{i-1}\}) \\ &\geq f(Q) - \sum_{i=1}^{|Q-S|} \Delta f(q_i|S) && \text{(submodularity)} \\ &> f(Q) - \epsilon \cdot f(S). \end{aligned}$$

The result is clear by rearranging the above inequality. □

Remark 1 Lemma 1 indicates that, at the termination of Algorithm 2, $f(S)$ gets close to $f(Q)$ if ϵ is small enough. This means that if the *mgv* of an element is less than $\frac{\epsilon}{r} d$, then this element can be considered negligible because it has very limited contribution to $f(S)$. This is the reason why the terminal threshold is set as $\frac{\epsilon}{r} d$.

Lemma 2 $\mathbb{E}[|K_u|] \leq Pr_{max}$ *where* $Pr_{max} = \max(pk, 1-p)$.

Proof There are three cases to analyse, depending on whether the current element u is considered at some point of iteration, i.e. $u \in C$, and whether u is already in Q at the beginning of the iteration in Algorithm 2. Note that the size of K_u is kept as small as possible.

i. If $u \notin C$ for whole iterations, $K_u = \emptyset$ and thus the expectation is obtained as:

$$\mathbb{E}[|K_u|] = 0.$$

ii. If $u \in C$ and $u \in Q$ at the beginning of the iteration, then $K_u = \emptyset$ for $u \in \mathcal{N}_s$ and $K_u = \{u\}$ for $u \notin \mathcal{N}_s$. Since u is sampled in $\mathcal{N}_s$ with probability p, the expectation is obtained as:

$$\mathbb{E}[|K_u|] = p \cdot |\emptyset| + (1-p)|\{u\}| = 1 - p.$$

iii. If $u \in C$ and $u \notin Q$ at the beginning of the iteration, then K_u contains at most k elements for $u \in \mathcal{N}_s$, and $K_u = \emptyset$ for $u \notin \mathcal{N}_s$. According to the property of k-extendible systems, if Q becomes dependent after adding u, then Q can remove at most

k elements to remain independence. If Q is still independent after adding u, then $K_u = \emptyset$. Therefore,

$$\mathbb{E}[|K_u|] \leq p \cdot k + (1-p)|\emptyset| = pk.$$

In summary, $\mathbb{E}[|K_u|] \leq \max(pk, 1-p)$. □

Lemma 3 $\mathbb{E}[f(S)] = \sum_{u \in \mathcal{N}} p\mathbb{E}[\Delta f(u|S_u)]$.

Proof Let us define a random variable $\mathcal{G}_u$ such that its value is equal to the increase of $f(S)$ when $u \in \mathcal{N}$ is considered, i.e.

$$f(S) = f(\emptyset) + \sum_{u \in \mathcal{N}} \mathcal{G}_u.$$

Note that since f is assumed to be normalised, $f(\emptyset) = 0$. Given the event $\mathcal{E}_u$ specifying all the decisions made before considering u, the conditional expectation of $\mathcal{G}_u$ is obtained as

$$\mathbb{E}[\mathcal{G}_u|\mathcal{E}_u] = \sum_{\mathcal{G}_u} P(\mathcal{G}_u|\mathcal{E}_u)\mathcal{G}_u.$$

Here, if u is sampled, $\mathcal{G}_u$ is equal to $\Delta f(u|S'_u)$ with the probability of $P(\mathcal{G}_u|\mathcal{E}_u) = p$, where S'_u is defined as S_u given the event $\mathcal{E}_u$. Note that if u is sampled but not in C, $\Delta f(u|S'_u)$ is defined as 0 by convention. Otherwise if u is not sampled, $\mathcal{G}_u$ is zero. Hence, the conditional expectation of $\mathcal{G}_u$ is:

$$\mathbb{E}[\mathcal{G}_u|\mathcal{E}_u] = p\Delta f(u|S'_u) = p\mathbb{E}[\Delta f(u|S_u)|\mathcal{E}_u].$$

By the law of total expectation, the expectation of $\mathcal{G}_u$ is obtained as:

$$\mathbb{E}[\mathcal{G}_u] = \mathbb{E}[\mathbb{E}[\mathcal{G}_u|\mathcal{E}_u)]] = \sum_{\mathcal{E}_u} P(\mathcal{E}_u)\mathbb{E}[\mathcal{G}_u|\mathcal{E}_u] = p\mathbb{E}[\Delta f(u|S_u)].$$

Hence, the expectation of $f(S)$ is obtained as:

$$\mathbb{E}[f(S)] = \sum_{u \in \mathcal{N}} p\mathbb{E}[\Delta f(u|S_u)].$$

□

Lemma 4 $\mathbb{E}[f(S)] > \frac{(1-\epsilon)p}{(1-\epsilon^2)p+Pr_{max}}\mathbb{E}[f(S \cup OPT)]$.

Proof In a certain iteration and given the current threshold θ, if $u \in C$ it implies that

$$\Delta f(u|S_u) \geq \theta. \tag{2}$$

While if an element $q \in K_u - S$ was not selected before this iteration, then

$$\Delta f(q|S_u) < \theta/(1-\epsilon). \tag{3}$$

Combining Eqs. (2) and (3) yields

$$\Delta f(u|S_u) > (1-\epsilon)\Delta f(q|S_u)\ \forall q \in K_u - S. \tag{4}$$

Additionally, any element can be removed from Q at most once. In other words, the element that is contained in K_u at one iteration is always different from those in other iterations when K_u is not empty. Therefore, the sets $\{K_u - S\}_{u\in\mathcal{N}}$ are disjoint. According to the definition and evolution of Q, Q can be expressed as

$$Q = (OPT - \cup_{u\in\mathcal{N}} K_u) \cup S = (S \cup OPT) - \cup_{u\in\mathcal{N}}(K_u - S). \tag{5}$$

Denote $\mathcal{N}$ as $\mathcal{N} = \{u_1, u_2, \cdots, u_{|\mathcal{N}|}\}$. Then we define Q_u^i as

$$Q_u^i \doteq (S \cup OPT) - \cup_{u\in\mathcal{N}_i}(K_u - S)$$

where $\mathcal{N}_i = \{u_1, \cdots, u_i\}$. Denote K_u and S_u corresponding to u_i in the i-th iteration as K_u^i and S_u^i, respectively. It is clear that $S_u^i \subseteq S \subseteq Q_u^i$. Using Eq. (5), one can have

$$\begin{aligned}
f(Q) &= f(S \cup OPT) - \sum_{i=1}^{|\mathcal{N}|} \Delta f(K_u^i - S|Q_u^i) \\
&\geq f(S \cup OPT) - \sum_{i=1}^{|\mathcal{N}|} \sum_{q\in K_u^i - S} \Delta f(q|S_u^i) && \text{(submodularity)} \\
&> f(S \cup OPT) - \sum_{u\in\mathcal{N}} |K_u - S| \frac{1}{1-\epsilon} \Delta f(u|S_u) && \text{(Eq. (4))} \\
&\geq f(S \cup OPT) - \sum_{u\in\mathcal{N}} |K_u| \frac{1}{1-\epsilon} \Delta f(u|S_u).
\end{aligned}$$

Taking expectation over $f(S)$ yields

$$\begin{aligned}
\mathbb{E}[f(S)] &> \frac{1}{1+\epsilon}\mathbb{E}[f(Q)] && \text{(Lemma 1)} \\
&> \frac{1}{1+\epsilon}\mathbb{E}[f(S \cup OPT)] - \frac{1}{(1+\epsilon)(1-\epsilon)} \cdot \mathbb{E}[|K_u|] \cdot \sum_{u\in\mathcal{N}} \mathbb{E}[\Delta f(u|S_u)] \\
&\geq \frac{1}{1+\epsilon}\mathbb{E}[f(S \cup OPT)] - \frac{1}{(1+\epsilon)(1-\epsilon)} \cdot Pr_{max} \cdot \sum_{u\in\mathcal{N}} \mathbb{E}[\Delta f(u|S_u)] && \text{(Lemma2)} \\
&= \frac{1}{1+\epsilon}\mathbb{E}[f(S \cup OPT)] - \frac{1}{(1+\epsilon)(1-\epsilon)} \cdot \frac{Pr_{max}}{p} \cdot \mathbb{E}[f(S)]. && \text{(Lemma3)}
\end{aligned}$$

The result is clear by rearranging the above inequality. □

Let us finish the proof of Theorem 1 in the following part of this section.

Proof (Theorem 1) Recall that, p is the sampling probability and $p \in (0, 1]$. Hence

$$Pr_{max} = \max(pk, 1-p) = \begin{cases} 1-p & \text{for } p \in (0, \frac{1}{1+k}] \\ pk & \text{for } p \in (\frac{1}{1+k}, 1]. \end{cases}$$

It is necessary to analyse the relationship between $f(S \cup OPT)$ and $f(OPT)$ with monotone and non-monotone submodular objective functions, respectively, to get the approximation guarantees for both cases.

- If f is monotone, then $f(S \cup OPT) \geq f(OPT)$. According to Lemma 4,

$$\begin{aligned}\mathbb{E}[f(S)] &> \frac{(1-\epsilon)p}{(1-\epsilon^2)p + Pr_{max}} \cdot \mathbb{E}[f(S \cup OPT)] \\ &\geq \frac{(1-\epsilon)p}{(1-\epsilon^2)p + Pr_{max}} \cdot f(OPT).\end{aligned}$$

When $p \in (0, \frac{1}{1+k}]$, it holds that

$$\begin{aligned}\mathbb{E}[f(S)] &> \frac{(1-\epsilon)p}{(1-\epsilon^2)p + 1 - p} \cdot f(OPT) \\ &> (p-\epsilon) \cdot f(OPT).\end{aligned}$$

When $p \in (\frac{1}{1+k}, 1]$, it holds that

$$\begin{aligned}\mathbb{E}[f(S)] &> \frac{(1-\epsilon)p}{(1-\epsilon^2)p + pk} \cdot f(OPT) \\ &> (\frac{1}{1+k} - \epsilon) \cdot f(OPT).\end{aligned}$$

- If f is non-monotone, let us define a new submodular and non-monotone function $h : 2^{\mathcal{N}} \to \mathbb{R}_{\geq 0}$ as $h(X) = f(X \cup OPT)\ \forall X \subseteq \mathcal{N}$. Since S contains each element with probability at most p and according to Claim 1, it is clear that

$$\mathbb{E}[f(S \cup OPT)] = \mathbb{E}[h(S)] \geq (1-p)h(\emptyset) = (1-p)f(OPT). \tag{6}$$

Combining Eq. (6) with Lemma 4 yields

$$\begin{aligned}\mathbb{E}[f(S)] &> \frac{(1-\epsilon)p}{(1-\epsilon^2)p + Pr_{max}} \cdot \mathbb{E}[f(S \cup OPT)] \\ &\geq \frac{(1-\epsilon)p(1-p)}{(1-\epsilon^2)p + Pr_{max}} \cdot f(OPT).\end{aligned}$$

When $p \in (0, \frac{1}{1+k}]$, it holds that

$$\begin{aligned}\mathbb{E}[f(S)] &> \frac{(1-\epsilon)p(1-p)}{(1-\epsilon^2)p + 1 - p} \cdot f(OPT) \\ &> [p(1-p) - \epsilon] \cdot f(OPT).\end{aligned}$$

When $p \in (\frac{1}{1+k}, 1]$, it holds that

$$\begin{aligned}\mathbb{E}[f(S)] &> \frac{(1-\epsilon)p(1-p)}{(1-\epsilon^2)p + pk} \cdot f(OPT) \\ &> (\frac{1}{1+k} - \epsilon)(1-p) \cdot f(OPT).\end{aligned}$$

In summary, if f is monotone, the expected approximation ratios are

$$\mathbb{E}[f(S)] > \begin{cases} (p-\epsilon) \cdot f(OPT) & \text{for } p \in (0, \frac{1}{1+k}] \\ (\frac{1}{1+k} - \epsilon) \cdot f(OPT) & \text{for } p \in (\frac{1}{1+k}, 1]. \end{cases} \tag{7}$$

If f is non-monotone, the expected approximation ratios are

$$\mathbb{E}[f(S)] > \begin{cases} [p(1-p)-\epsilon]\cdot f(OPT) & \text{for } p \in (0, \frac{1}{1+k}] \\ (\frac{1}{1+k}-\epsilon)(1-p)\cdot f(OPT) & \text{for } p \in (\frac{1}{1+k}, 1]. \end{cases} \tag{8}$$

Eqs. (7) and (8) show that, for $p \in (\frac{1}{1+k}, 1]$, the expected approximation ratio becomes stagnated in the monotone case and decreasing in the non-monotone case. Moreover, the computational complexity increases as the sampling probability gets larger. On the other side, for $p \in (0, \frac{1}{1+k}]$, the sampling probability provides adjustment capability for the trade-off between the approximation ratio and computational complexity. As the probability increases for $p \in (0, \frac{1}{1+k}]$, the expected approximation ratios improve for both monotone and non-monotone cases, but the computational complexity also increases.

Recall that the theoretical time complexity is $O\big(\frac{pn}{\epsilon}\ln\frac{r}{\epsilon}\big)$. The impact of ϵ on the solution quality and time complexity is more desirable than that of p. Therefore, this work fixes the sampling probability as $p = \frac{1}{1+k}$ and leave ϵ as an adjustable designing parameter for the trade-off of solution quality versus time complexity. According to Eqs. (7) and (8), the best expected approximation ratios can be readily obtained, when $p = \frac{1}{1+k}$, as:

$$\mathbb{E}[f(S)] > \begin{cases} (\frac{1}{1+k}-\epsilon)\cdot f(OPT) & \text{if } f \text{ is monotone} \\ [\frac{k}{(1+k)^2}-\epsilon]\cdot f(OPT) & \text{if } f \text{ is non-monotone}. \end{cases}$$

□

Experiments

This section testifies the proposed algorithm SDTG through experiments using a real database and compares its performance with that of Greedy [19] and Sample [21]. For a fair comparison, this section uses the basic versions of these algorithms without integrating the Lazy strategy [38]. Note that the performance of Sample and FANTOM [5] has already been compared in [21].

Experimental setup

The database used in the experiments is MovieLens 20M [23]. This database contains 20 million ratings and 465,000 tag applications applied to 27,000 movies by 138,000 users. Movies in the database are classified into 19 genres, such as Action, Comedy, Drama, etc. Besides, each movie is also scored according to the relevance with 1128 genome tags forming 12 million relevance scores in total.

The objective of the movie recommendation system in the experiments is to select a shortlist of movies that are representative yet diverse for users based on their favourite movie genres. The objective function is introduced from [5, 21]. Let $\mathcal{N}$ be the set of all movies and G be the set of all movie genres. Denote $\mathcal{N}(g)$ as the set of all movies that belong to the movie genre $g \in G$. Denote $G(i)$ as the set of genres that the movie i belongs to. Note that one movie can belong to different genres, hence $|G(i)| \geq 1$. Let s_{ij} represent the similarity between movie i and movie j. Denote G_μ as the set of all movie genres that the user μ likes, $G_\mu \subseteq G$. The movies that can be considered by the user μ is

contained in the set $\mathcal{N}_\mu = \cup_{g \in G_\mu} \mathcal{N}(g)$. The objective function of movie recommendation for user μ is given by

$$f_\mu(S) = \sum_{i \in S} \sum_{j \in \mathcal{N}_\mu} s_{ij} - \lambda \sum_{i \in S} \sum_{j \in S} s_{ij} \tag{9}$$

where $\lambda \in [0, 1]$ is the penalty parameter for the similarity between movies within the recommendation list S. The objective function Eq. (9) is non-negative, non-monotone, and submodular. The first term of Eq. (9) reflects the representativeness of the selected movies, and the second term helps to increase diversity. It is desired to achieve high objective function value with low computational complexity.

The similarity value between movie i and movie j can be calculated based on the Euclidean distance of relevance scores

$$s_{ij} = \frac{1}{\sqrt{\sum_{t=1}^{N_t} (\gamma_t^i - \gamma_t^j)^2}}$$

where $N_t = 1128$ is the number of all genome tags, γ_t^i and γ_t^j are the relevance scores in terms of the tag t for movie i and movie j, respectively. The calculation of the similarity map took around 35 days on Cranfield HPC—Delta,[1] using 128 CPUs with parallel computing.

The constraints of the movie recommendation system come from the upper limits of the number of movies in total and in each movie genre. The first constraint is an upper limit m on the total number of movies in the movie recommendation list for the user. The second one is an upper limit m_g (named as a *genre limit*) on the number of movies that belong to the movie genre g. According to [21], the movie recommendation system is subject to a $|G_\mu|$-extendible system constraint.

In the experiments, suppose that the user's favourite movie genres are Action, Adventure, and Sci-Fi. Then, the constraint of the movie recommendation system is a 3-extendible system constraint. Movies with ids less than 30,000 are within consideration since not all movies have genome scores in the database. Set the upper limit on the total number of movies as $m = 15$, and the genre limit as varying numbers from 1 to 6. Set the sampling probability for Sample and SDTG as $p = 0.25$, and the threshold decreasing parameter for SDTG as $\epsilon = 0.2$. Set the penalty parameter as $\lambda = 0.8$. Denoted Max Sample (4) and Max SDTG (4) as the best selections from 4 rounds of Sample and SDTG, respectively. The results of Sample and SDTG are based on 100 rounds of these two algorithms. The running time for these algorithms is measured as the number of objective function evaluations which is independent on the computer conditions. Note that, the experimental results for Sample and SDTG vary somehow each time as the algorithms are related to random sampling.

[1] Please refer to https://www.cranfield.ac.uk/study/it-services for details about Delta. Accessed 15 Dec 2020.

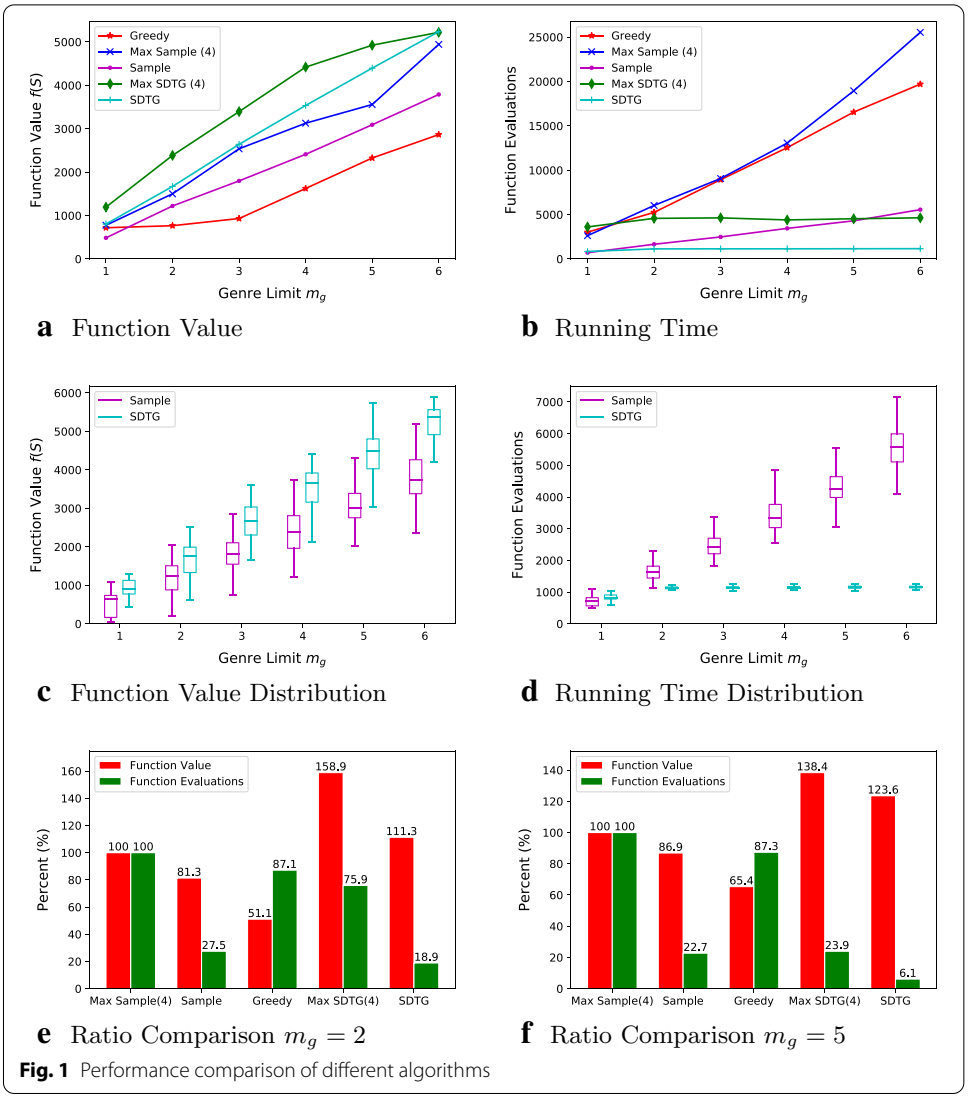


Fig. 1 Performance comparison of different algorithms

Results

The performance of SDTG is compared with that of benchmark algorithms in terms of both function values and running time in Fig. 1. It is clear from Fig. 1a that, on average, Sample and SDTG related algorithms outperform Greedy in terms of solution quality. The quality of solutions provided by SDTG is better than that of Sample, although SDTG has a slightly worse theoretical approximation guarantee than Sample does. Overall, Max SDTG (4) achieves the highest function value. Figure 1b shows the number of function evaluations consumed by different algorithms. Four rounds of Sample requires the largest number of function evaluations when $m_g \geq 2$. Relatively, Greedy requires a bit fewer function evaluations than Max Sample (4) does. But four rounds of SDTG requires significantly fewer evaluations. Overall, Greedy and Sample-related algorithms consume increasing numbers of function evaluations as m_g goes up. However, the numbers of function evaluations of the SDTG-related algorithms almost stay constant when $m_g \geq 2$. When $m_g = 6$, four rounds of SDTG is even faster than one round of Sample.

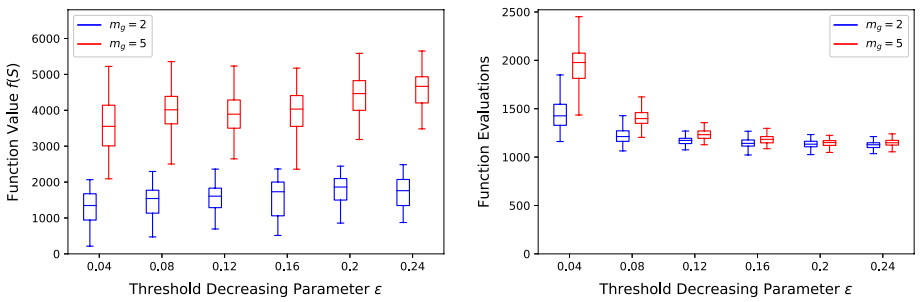
Table 1 Movies recommended by different algorithms, $m_g = 2$

Algorithm	Movie id	Genres
Greedy	2367	Adventure, Fantasy, Romance, Sci-Fi, Thriller
	26513	Action, Adventure, Comedy, Sci-Fi
	4629	Action, Crime, Thriller
Max Sample (4)	4629	Action, Crime, Thriller
	736	Action, Adventure, Romance, Thriller
	6106	Adventure
	4545	Comedy, Sci-Fi
	4738	Romance, Sci-Fi
Max SDTG (4)	4369	Action, Crime, Thriller
	42	Action, Crime, Drama
	146	Adventure, Children
	231	Adventure, Comedy
	1965	Comedy, Sci-Fi
	2656	Horror, Sci-Fi

42 146 231 736 1965 2367 2656 4369 4545 4629 4738 6106 26513

Figure 1c, d illustrate the distribution of function value and running time for 100 rounds of Sample and SDTG algorithms. Recall that, Max Sample (4) and Max SDTG (4) represent the maximum values achieved by four rounds of Sample and SDTG, respectively. And Greedy is a deterministic algorithm. Therefore, these three items do not appear in Fig. 1c, d that are for demonstrating the distribution resulted from random sampling. Overall, the function value distribution of SDTG has similar spreads with Sample's, but SDTG achieves higher median values than Sample does. In terms of running time, SDTG has significantly smaller spreads and lower median values than Sample does. The comparison between Sample and SDTG indicates that SDTG not only achieves better function values but also is faster and more reliable.

Figure 1e, f demonstrate the ratio comparison of the solution quality and running time of different algorithms. The performance of Max Sample (4) is set as a baseline for other algorithms in comparison. When $m_g = 2$, Max SDTG (4) achieves a significantly better function value but consumes fewer function evaluations than Max Sample (4) does. While $m_g = 5$, Max SDTG (4) achieves a much better function value (38.4% higher) and consumes a dramatically smaller number of function evaluations (76.1% fewer). On average, SDTG finds better solutions but only consumes 6.1% of function evaluations compared with Max Sample (4). In both cases, Greedy is the least competitive one among all algorithms because it achieves the worst function values and requires the second largest number of function evaluations. SDTG provides high-quality solutions yet consumes the fewest function evaluations, which is of great advantage when handling large-scale datasets.



a Function Value Distribution **b** Running Time Distribution

Fig. 2 The effect of ϵ on function value and running time

Discussion

The reason why Greedy performs poorly in terms of solution quality is that it greedily selects the best element during each iteration heading to bad local optima. On the other side, with the help of the sampling process, Sample and SDTG related algorithms are able to avoid those elements that can get the algorithms trapped in bad local optima. The threshold in SDTG can further help the algorithm to avoid those local optima. This is why SDTG practically outperforms Sample in terms of solution quality. Table 1 explains the reason in detail. According to the definition of the genre limit constraint, at most two movies can be selected from each genre of Adventure, Action, and Sci-Fi when $m_g = 2$. The maximum number of movies without violating the aforementioned constraint is six. Greedy only recommends three movies and reaches the upper genre limit. However, Max Sample (4) and Max SDTG (4) are able to recommend five and six movies, respectively, which better fit the objective of the movie recommendation system.

The reason why Greedy performs poorly in terms of running time is that it has to calculate the *mgv*s of all remaining elements given the current selection to find the best one. Sample is faster than Greedy because it only considers a small portion of the ground set, although it also needs to evaluate all remaining elements in the sample set. Different from Sample, SDTG can stop evaluating once it finds one qualified element and adds this element to the selection set immediately. This means that SDTG does not have to evaluate all the remaining elements in the sample set in order to select an extra element. Therefore, SDTG consumes fewer function evaluations than Sample does on average. In addition, the running time of Sample is highly dependent on the size of the sample set because it needs to evaluate all elements in the sample set. In contrast, SDTG can usually find a qualified element from the front positions of the sample set and stop evaluating. Therefore, the running time of SDTG is less related to the size of the sample set compared with Sample's. This is the reason why the spread of running time distribution of SDTG is smaller than Sample's.

Trade-off of solution quality vs. running time

This section also examines the impact of the threshold parameter ϵ on solution quality and running time. This will help us to choose a desirable value of ϵ and to have a

deeper comprehension of SDTG. The value of ϵ varies from 0.04 to 0.24 with a step of 0.04. Two cases are checked where m_g equals to 2 and 5, respectively. Other settings are as same as previous ones. We run 100 rounds of SDTG and record the function values and the number of function evaluations in each round.

Figure 2 demonstrates the experimental results with varying values of the threshold decreasing parameter. The distributions of function value and running time are illustrated in Fig. 2a, b, respectively. Figure 2a shows that the impact of changing ϵ on function values is not significant. Function values fluctuate slightly when $\epsilon \geq 0.08$. However, the solution quality for both $m_g = 2$ and $m_g = 5$ is obviously worse when ϵ equals to 0.04 than that with larger values of ϵ. This is because the threshold decreases very slowly with an extremely small ϵ. In this case, the *mgv* of the element selected by SDTG in each iteration is very close to the largest one. As mentioned before, the decreasing threshold can also help SDTG to avoid local optima. An extremely small ϵ makes SDTG close to Sample, which weakens the advantage of the decreasing threshold. Figure 2b shows that the median values of running time decrease obviously as ϵ increases. The spreads of running time also become smaller as ϵ goes up. The reason is that the threshold decreases faster with a larger ϵ. When evaluating the *mgv*s of the remaining elements one by one, SDTG can find a qualified element more quickly with a smaller threshold. The running time of SDTG also becomes less dependent on the size of the sample set.

Conclusions

This paper has presented an efficient algorithm, Sample Decreasing Threshold Greedy (SDTG), to deal with big data summarisation problems. The proposed algorithm achieves an expected approximation ratio of $\frac{k}{(1+k)^2} - \epsilon$ for maximising general non-monotone submodular objective functions subject to k-extendible system constraints with only $O\left(\frac{n}{(1+k)\epsilon} \ln \frac{r}{\epsilon}\right)$ value oracle calls. The performance of SDTG is testified and compared with that of benchmark algorithms through experiments with a movie recommendation system based on a widely-used movie information database. The experimental results indicate that the proposed algorithm has great application potentials in large-scale discrete optimisation problems where the sizes of datasets are enormous such as the applications of machine learning and big data science. We believe that our results are also instrumental for the personalised recommendation systems on internet platforms, like Netflix, YouTube, and Amazon, etc. SDTG can be further accelerated by adapting the Lazy Greedy strategy [38]. A future research direction could also be accelerating the proposed algorithm by combining distributed computing.

Abbreviations

SDTG: Sample decreasing threshold greedy; *mgv*: Marginal gain value; OPT: Optimal solution; Max Sample (4): Run 4 rounds of Sample and get the maximum function value; Max SDTG (4): Run 4 rounds of SDTG and get the maximum function value.

Acknowledgements

The authors thank the Cranfield IT Department team for helping with the Cranfield HPC—Delta operations.

Authors' contributions

TL contributed to the algorithm design and analysis, experiments, and manuscript drafting. HS contributed to the theoretical and experimental analysis, manuscript drafting. AT helped to arrange the resources required by the experiments. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets generated during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 4 November 2020 Accepted: 16 January 2021
Published online: 09 February 2021

References

1. Jin X, Wah BW, Cheng X, Wang Y. Significance and challenges of big data research. Big Data Res. 2015;2(2):59–64.
2. Mirzasoleiman B. Big data summarization using submodular functions. Doctoral dissertation, ETH Zurich; 2017.
3. Tschiatschek S, Djolonga J, Krause A. Learning probabilistic submodular diversity models via noise contrastive estimation. In: Proceedings of the 19th international conference on artificial intelligence and statistics (AISTATS); 2016. p. 770–9.
4. Yu Q, Xu EL, Cui S. Submodular maximization with multi-knapsack constraints and its applications in scientific literature recommendations. In: 2016 IEEE global conference on signal and information processing (GlobalSIP); 2016. p. 1295–9.
5. Mirzasoleiman B, Badanidiyuru A, Karbasi A. Fast constrained submodular maximization: personalized data summarization. In: Proceedings of the 33rd international conference on machine learning (ICML). vol. 48; 2016. p. 1358–67.
6. Mirzasoleiman B, Karbasi A, Krause A. Deletion-robust submodular maximization: data summarization with the right to be forgotten. In: Proceedings of the 34th international conference on machine learning (ICML). vol. 70; 2017. p. 2449–58.
7. Mirzasoleiman B, Karbasi A, Sarkar R, Krause A. Distributed submodular maximization: identifying representative elements in massive data. In: Advances in neural information processing systems (NIPS); 2013. p. 2049–57.
8. Mirzasoleiman B, Karbasi A, Sarkar R, Krause A. Distributed submodular maximization. J Mach Learn Res. 2016;17(1):8330–733.
9. Norouzi-Fard A, Tarnawski J, Mitrović S, Zandieh A, Mousavifar A, Svensson O. Beyond 1/2-approximation for submodular maximization on massive data streams. In: Proceedings of the 35th international conference on machine learning (ICML); 2018. p. 3829–38.
10. Balkanski E, Mirzasoleiman B, Krause A, Singer Y. Learning sparse combinatorial representations via two-stage submodular maximization. In: Proceedings of the 33rd international conference on machine learning (ICML); 2016. p. 2207–16.
11. Lavania C, Bilmes J. Auto-summarization: a step towards unsupervised learning of a submodular mixture. In: Proceedings of the 2019 SIAM international conference on data mining (sDM). SIAM; 2019. p. 396–404.
12. Badanidiyuru A, Mirzasoleiman B, Karbasi A, Krause A. Streaming submodular maximization: massive data summarization on the fly. In: 20th ACM SIGKDD international conference on knowledge discovery and data mining (KDD). New York: ACM; 2014. p. 671–80.
13. Balkanski E, Breuer A, Singer Y. Non-monotone submodular maximization in exponentially fewer iterations. In: 32nd conference on neural information processing systems (NeurIPS 2018); 2018. p. 2353–64.
14. Mitrovic M, Kazemi E, Zadimoghaddam M, Karbasi A. Data summarization at scale: a two-stage submodular approach. In: Proceedings of the 35th international conference on machine learning (ICML). vol. 80. PMLR; 2018. p. 3596–605.
15. Mirzasoleiman B, Karbasi A, Badanidiyuru A, Krause A. Distributed submodular cover: succinctly summarizing massive data. In: Advances in neural information processing systems (NIPS); 2015. p. 2881–9.
16. Xu J, Mukherjee L, Li Y, Warner J, Rehg JM, Singh V. Gaze-enabled egocentric video summarization via constrained submodular maximization. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR); 2015. p. 2235–44.
17. Gygli M, Grabner H, Van Gool L. Video summarization by learning submodular mixtures of objectives. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR); 2015. p. 3090–8.
18. Krause A, Guestrin C. Near-optimal observation selection using submodular functions. In: Proceedings of the 22nd national conference on artificial intelligence. vol. 2. Palo Alto: AAAI Press; 2007. p. 1650–4.
19. Nemhauser GL, Wolsey LA, Fisher ML. An analysis of approximations for maximizing submodular set functions—I. Math Program. 1978;14(1):265–94.
20. Mestre J. Greedy in approximation algorithms. In: European symposium on algorithms (ESA). New York: Springer; 2006. p. 528–39.
21. Feldman M, Harshaw C, Karbasi A. Greed is good: near-optimal submodular maximization via greedy optimization. In: Proceedings of the 2017 conference on learning theory (COLT). vol. 65. PMLR; 2017. p. 1–27.
22. Badanidiyuru A, Vondrák J. Fast algorithms for maximizing submodular functions. In: Proceedings of the 25th annual ACM-SIAM symposium on discrete algorithms (SODA). SIAM; 2014. p. 1497–514.
23. Harper FM, Konstan JA. The movielens datasets: history and context. ACM Trans Interac Intell Syst (TIIS). 2016;5(4):1–19.
24. Mirzasoleiman B, Badanidiyuru A, Karbasi A, Vondrák J, Krause A. Lazier than lazy greedy. In: 29th AAAI conference on artificial intelligence. Palo Alto: AAAI Press; 2015. p. 1812–8.
25. Buchbinder N, Feldman M, Schwartz R. Comparing apples and oranges: query trade-off in submodular maximization. Math Oper Res. 2016;42(2):308–29.
26. Breuer A, Balkanski E, Singer Y. The FAST algorithm for submodular maximization. In: International conference on machine learning. PMLR; 2020. p. 1134–43.

27. Nemhauser GL, Wolsey LA. Best algorithms for approximating the maximum of a submodular set function. Math Oper Res. 1978;3(3):177–88.
28. Calinescu G, Chekuri C, Pál M, Vondrák J. Maximizing a monotone submodular function subject to a matroid constraint. SIAM J Comput. 2011;40(6):1740–66.
29. Feldman M, Naor J, Schwartz RA, unified continuous greedy algorithm for submodular maximization. In: 2011 IEEE 52nd annual symposium on foundations of computer science (FOCS). New York: IEEE. 2011. p. 570–9.
30. Amanatidis G, Fusco F, Lazos P, Leonardi S, Reiffenhäuser R. Fast adaptive non-monotone submodular maximization subject to a knapsack constraint. In: Advances in Neural Information Processing Systems. 2020;33.
31. Segui-Gasco P, Shin HS. Fast non-monotone submodular maximisation subject to a matroid constraint. arXiv preprint arXiv:170306053. 2017.
32. Gupta A, Roth A, Schoenebeck G, Talwar K. Constrained non-monotone submodular maximization: offline and secretary algorithms. In: International workshop on internet and network economics (WINE). New York: Springer; 2010. p. 246–57.
33. Li T, Shin HS, Tsourdos A. Fast submodular maximization subject to k-extendible system constraints. arXiv preprint arXiv :181107673v1. 2018.
34. Shin HS, Li T, Segui-Gasco P. Sample greedy based task allocation for multiple robot systems. arXiv preprint arXiv:19010 3258. 2019.
35. Li T, Shin HS, Tsourdos A. Threshold greedy based task allocation for multiple robot operations. arXiv preprint arXiv:19090 1239. 2019.
36. Krause A, Golovin D. Submodular function maximization. Tractability. 2014;3:71–104.
37. Buchbinder N, Feldman M, Naor JS, Schwartz R. Submodular maximization with cardinality constraints. In: Proceedings of the 25th annual ACM-SIAM symposium on discrete algorithms (SODA). SIAM; 2014. p. 1433–52.
38. Minoux M. Accelerated greedy algorithms for maximizing submodular set functions. Optimization techniques. Berlin: Springer; 1978. p. 234–243.

Publisher's Note

2021-02-09

A sample decreasing threshold þÿ g r e e d y b a s e d a l g o r i t h m f o r b i

Li, Teng

Springer

þÿ L i T , S h i n H - S & T s o u r d o s A (2 0 2 1) A s a m p l e d e c r e a s i n g t h r e s h o l d g r e e d
big data summarisation, Journal of Big Data, Volume 8, 2021, Article number 30
https://doi.org/10.1186/s40537-021-00416-y