Course : Data mining
Topic : Rank aggregation

Aristides Gionis
Aalto University
Department of Computer Science

visiting in Sapienza University of Rome
fall 2016

reading

Cynthia Dwork, Ravi Kumar, Moni Naor, D. Sivakumar:
Rank aggregation methods for the web.
WWW 2001

(optional)
Nir Ailon, Moses Charikar, Alantha Newman:
Aggregating inconsistent information: Ranking and clustering.
JACM 55(5), 2008

rank aggregation and voting

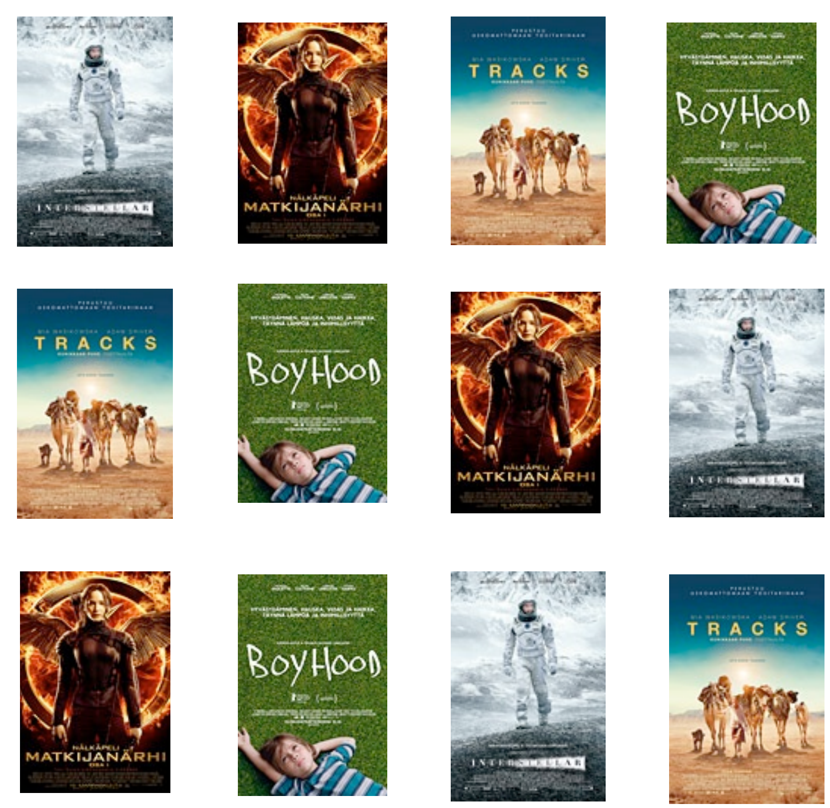
how can multiple agents aggregate their preferences and make a consensus decision?

example : three friends want to go to the cinema

Luca :

Stefano :

Aris :



which movie should they choose?

what are good properties for a voting system?

question considered by marquis de Condorcet (1743-1794)

French philosopher, mathematician and political scientist

proposed a criterion that voting systems should satisfy

known as the Condorcet criterion

what are good properties for a voting system

the Condorcet criterion

if item i defeats every other item in a pairwise majority vote, then i should be ranked first

extended Condorcet criterion

if all items in a set X defeat in pairwise comparisons all items in the set Y then the items in X should be ranked above those in Y

not all voting systems satisfy the Condorcet criterion!

the Borda count voting system

proposed by Jean-Charles de Borda (1733-1799)

French mathematician, physicist, political scientist, and sailor

very popular and widely-used system

the Borda count voting system

in each preference list, assign to item i number of points equal to the number of item it defeats

first position gets n-1 points, second n-2, ..., last 0 points

the total weight of i is the number of points it accumulates from all preference lists

order items in decreasing weight

Borda count satisfies a number of desirable properties, but not the Condorcet criterion

more recent attempts to design axiomatic voting systems

objective :

construct a voting system that satisfies a set of natural axioms

Kenneth Arrow, PhD thesis, 1963

Nobel prize in economics, 1972, for general economics equilibrium theory and welfare theory

Arrow’s axioms

non-dictatorship : the preferences of an individual should not become the group ranking without considering the preferences of others

unanimity (or Pareto optimality) : if every individual prefers one choice to another, then the group ranking should do the same

freedom from irrelevant alternatives : if a choice is removed, then the others' order should not change

impossibility of voting

Arrow's theorem :

it is impossible to construct a voting system that satisfies the previous set of three axioms

impossibility of voting
Arrow's axioms

freedom from irrelevant alternatives : if a choice is removed, then the others' order should not change

heavily disputed axiom

Borda count violates this axiom

still..

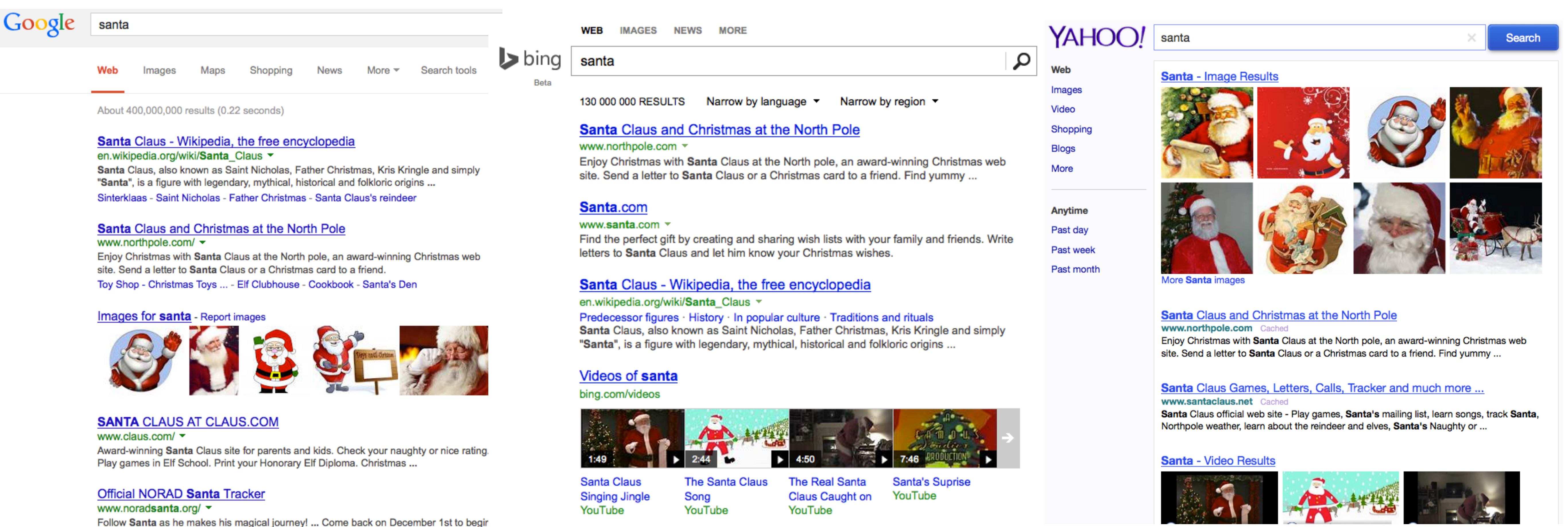
despite theoretical impossibility, the problem appears in practice and needs to be addressed

- selecting representatives in elections
- meta-search engines

meta-search engines

aggregate rankings from different search engines

- obtain better results than any individual one
- robust to spam



the rank-aggregation problem

input

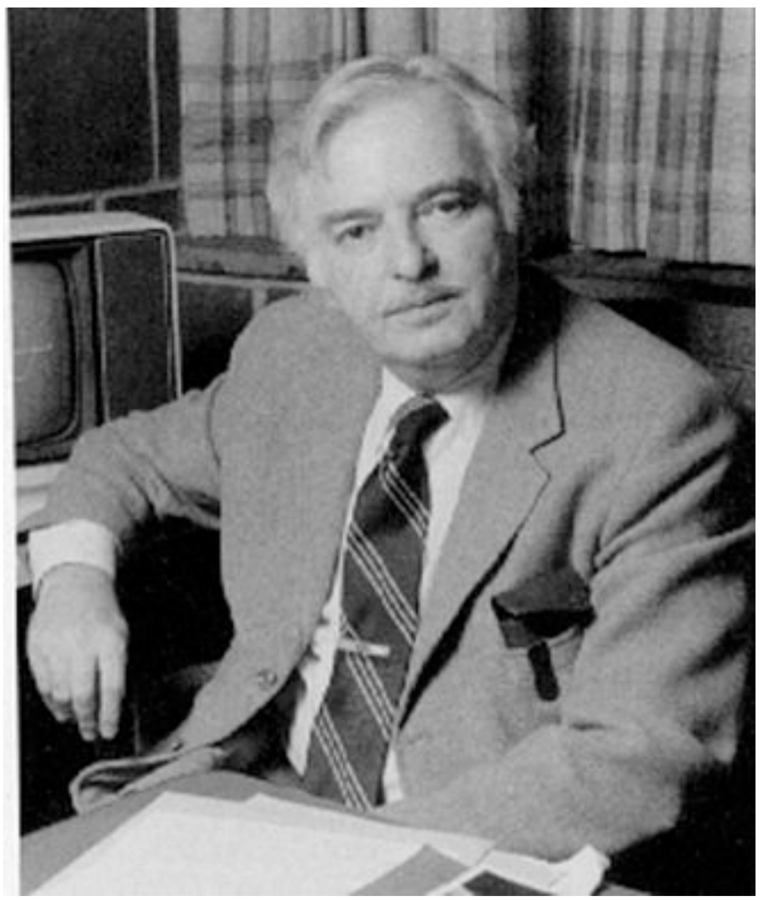
n items (movies, candidates, urls)

k preference lists (orderings) on the items

goal

find a single preference list that respects / agrees as much as possible with the input preference lists

Kemeny optimal aggregation



John Kemeny (1926-1992)

Hungarian-American mathematician and computer scientist

provided a specific formulation of the rank-aggregation problem

(also invented BASIC)

Kemeny optimal aggregation

input

n items (movies, candidates, urls)

k preference lists (orderings) on the items

goal

find a single preference list that minimizes the total number of out-of-order pairs

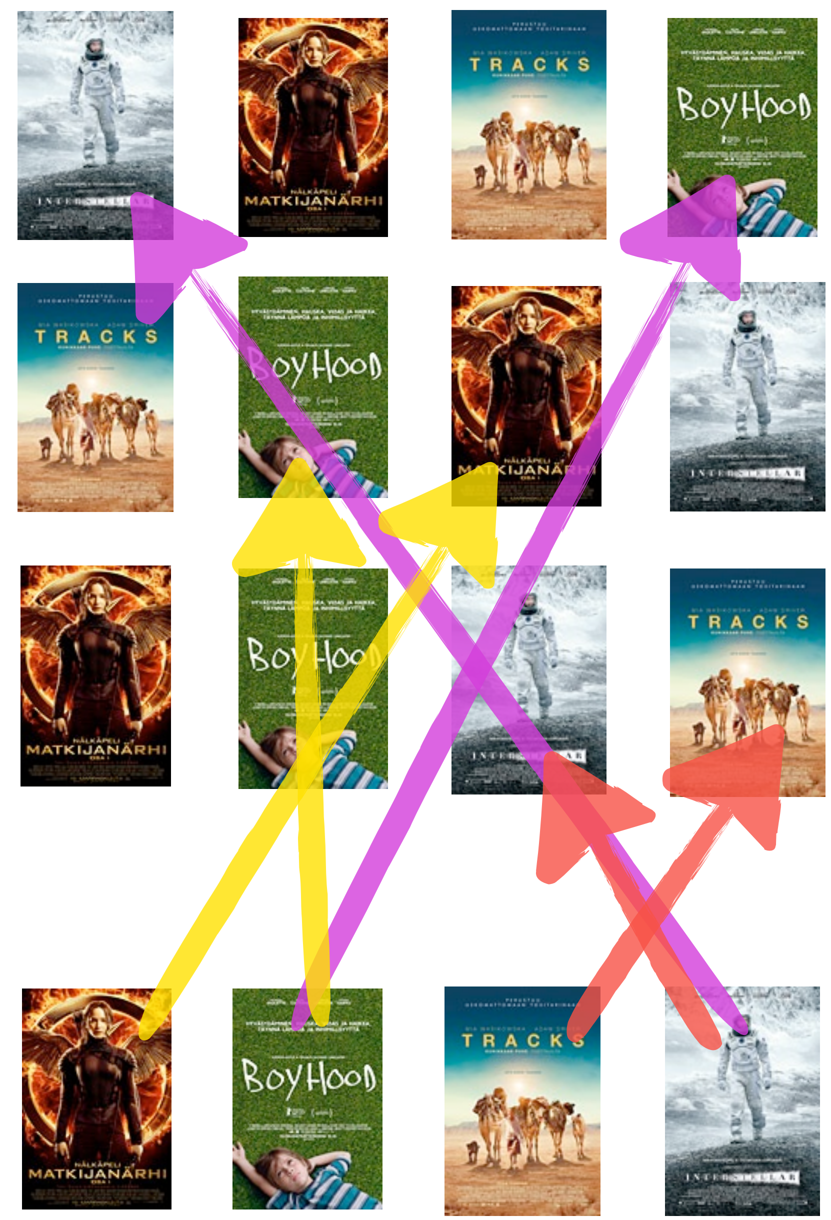
Kemeny optimal aggregation

Luca :

Stefano :

Aris :

aggregation :



preference lists

set of items U

assume n items

a preference list is a bijection (1-to-1 function) from U to {1,...,n}

for a preference list σ and item i in U denote by σ(i) the rank (order) of i in σ

preference lists can be:

full, partial, top-d

distances between preference lists

consider preference lists σ and τ over the same set of items U

how similar are σ and τ?

define a distance function

Spearman footrule distance

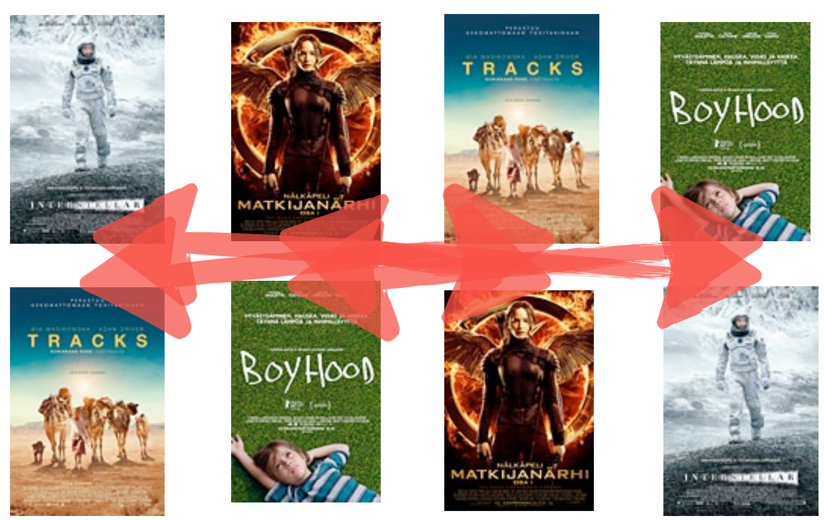
given two lists σ and τ over U, the Spearman footrule distance is defined as

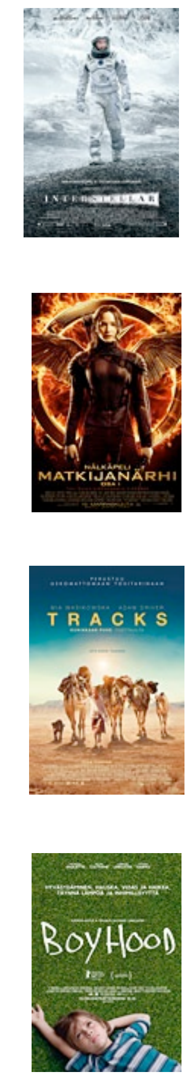
$$F(\sigma,\tau) = \sum_{i \in U} |\sigma(i) - \tau(i)|$$

Spearman footrule distance example

Luca :

Stefano :





3

1

2

2

F(Luca, Stefano) = 8

Kendall-tau distance

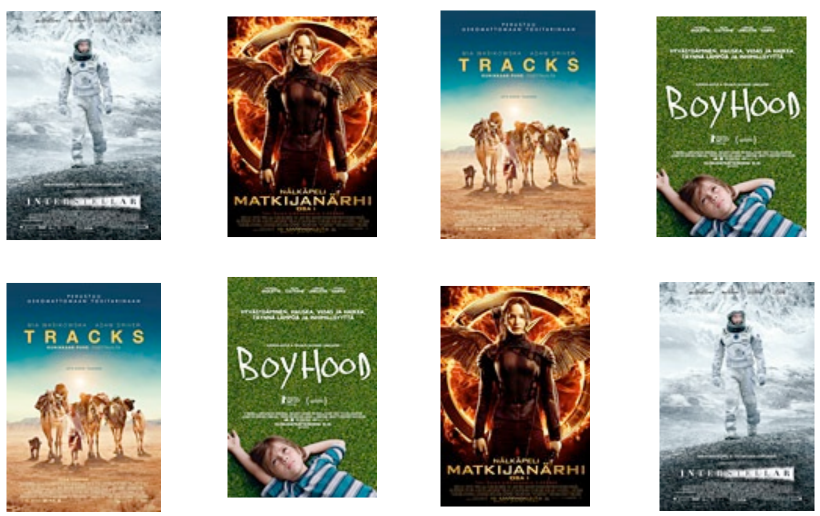
given two lists σ and τ over U, the Kendall-tau distance is the number of pair-wise disagreements

$$K(\sigma,\tau) = |\{(i,j) \text{ such that } \sigma(i) < \sigma(j) \text{ but } \tau(i) > \tau(j)\}|$$

Kendall-tau distance example

Luca :

Stefano :



K(Luca, Stefano) = 5

properties of Spearman footrule and Kendall-tau distances

are they metric?

definitions for full preference lists

what about partial lists?

the two distances F and K are related

for any two full preference lists:

$$K(\sigma,\tau) \leq F(\sigma,\tau) \leq 2K(\sigma,\tau)$$

the rank-aggregation problem

input

set U of n items

k preference lists $\tau_1,\ldots,\tau_k$

a distance function D between preference lists

(e.g., F or K)

goal

find preference list τ_0 that minimizes total disagreement

$$D(\tau_0,\tau_1\ldots\tau_k) = \sum_{i=1\ldots k} D(\tau_0,\tau_i)$$

when $D=K$, this is Kemeny optimal aggregation

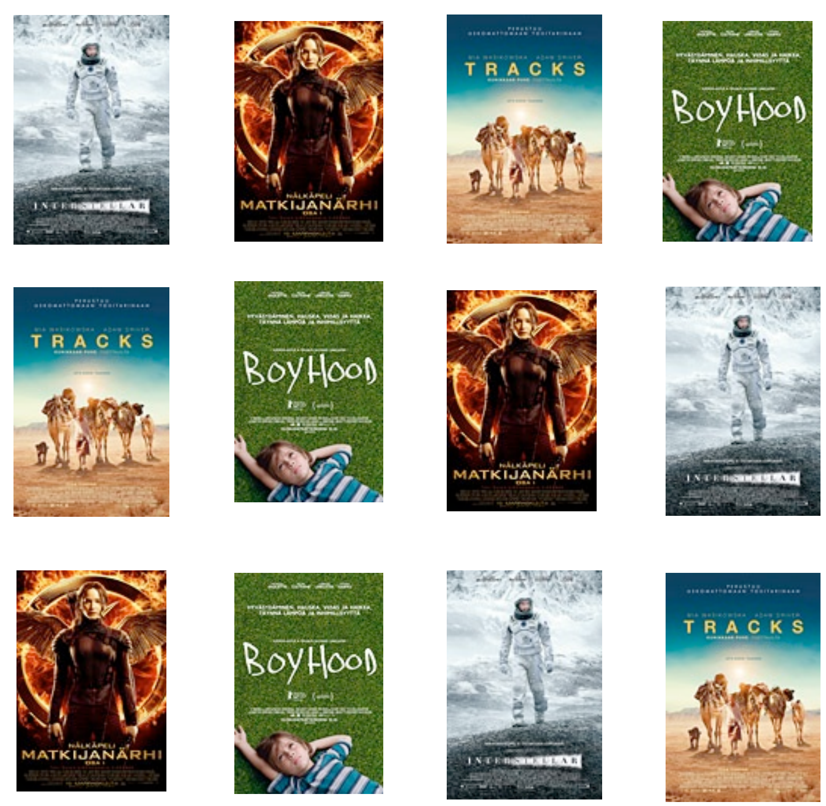
rank-aggregation with Spearman footrule distance

when distance is F the rank aggregation problem can be solved in polynomial time

Luca :

Stefano :

Aris :



rank-aggregation with Kendall-tau distance

when distance is K and $k \geq 4$ the rank aggregation problem is NP-hard!

but optimal preference list with Spearman footrule distance gives factor 2 approximation

τ_F : optimal list according to Spearman footrule

τ_0 : optimal list according to Kendall-tau

$$K(\tau_F,\tau_1 \ldots \tau_k) \leq F(\tau_F,\tau_1 \ldots \tau_k) \leq F(\tau_0,\tau_1 \ldots \tau_k) \leq 2K(\tau_0,\tau_1 \ldots \tau_k)$$

rank-aggregation with Kendall-tau distance

any other way to get a factor-2 approximation?

1-median problem in a metric space

algorithm : pick-the-best

try each one of $\tau_1,...,\tau_k$ as a potential solution and pick the best

algorithm pick-the-best is a factor 2 approximation

assume optimal solution τ_0

assume algorithm picked τ_j

assume τ_x is closest to τ_0 among all $\tau_1,\ldots,\tau_k$

$$D(\tau_j,\tau_1\ldots\tau_k) \leq D(\tau_x,\tau_1\ldots\tau_k)$$

$$= \sum_{i=1\ldots k} D(\tau_x,\tau_i)$$

$$\leq \sum_{i=1\ldots k} (D(\tau_x,\tau_0) + D(\tau_0,\tau_i))$$

$$= \sum_{i=1\ldots k} D(\tau_x,\tau_0) + \sum_{i=1\ldots k} D(\tau_0,\tau_i)$$

$$\leq \sum_{i=1\ldots k} D(\tau_0,\tau_i) + \sum_{i=1\ldots k} D(\tau_0,\tau_i) = 2\, D(\tau_0,\tau_1\ldots\tau_k)$$

yet another algorithm KwikSort [Ailon et al]

inspired by QuickSort

view data as a tournament over items in U

tournament: complete directed graph

for each pair i and j in U,

if the majority of preference lists prefer i over j put a directed edge from i to j

the KwikSort algorithm

pick a random element i in U

put at the left L all items that point to i

put at the right R all items that i points to

recurse on L and R

KwikSort gives a factor 3 approximation

but...

...taking the best of pick-the-best and KwikSort gives a factor 6/5 approximation!

Kemeny optimality and Condorcet criterion

Kemeny optimal aggregation satisfies the Condorcet criterion

but it is NP-hard to compute

can we have any other aggregation system that satisfies the Condorcet criterion?

locally Kemeny optimal aggregation

a ranking τ is locally Kemeny optimal if there is no bubble-sort swap of two consecutively placed items that produces a ranking τ' such that

$$K(\tau', \tau_1 \ldots \tau_k) \leq K(\tau, \tau_1 \ldots \tau_k)$$

locally Kemeny optimal is not necessarily Kemeny optimal

locally Kemeny optimal aggregation

locally Kemeny optimal aggregation can be computed in polynomial time

- proceed iteratively: in each iteration insert item i in the bottom of the list
- bubble it up until there is item j such that the majority places j over i

locally Kemeny optimal aggregation satisfies the Condorcet and extended Condorcet criterion

can be applied as post-processing to any rank aggregation system