

Data Mining

Web Advertising

Johannes Zschache
Wintersemester 2019

Abteilung Datenbanken, Universität Leipzig
<http://dbs.uni-leipzig.de>

Übersicht

Hochdimensionale Daten

Clustering

Dimensions-
reduktion

Empfehlungs-
systeme

Assoziations-
regeln

Locality Sensitive
Hashing

Supervised ML

Graphdaten

Community
Detection

PageRank

Web Spam

Datenströme

Windowing

Filtern

Momente

Web Advertising

Inhaltsverzeichnis

- **Einführung**
- **Greedy Matching Algorithmus**
- **Balance Matching Algorithmus**

Literatur: Kapitel 8 aus „Mining of Massive Datasets“: <http://www.mmds.org>

Werbung auf Webseiten

The screenshot shows the LEO dictionary interface. At the top, there is a navigation bar with 'Home', 'Dictionary', 'Forums', 'Trainer', and 'Courses'. Below this is the LEO logo and a search bar containing 'Web advertising'. A search result for 'Web advertising' is displayed, showing 814,442 entries and 7,617,493 queries today. The results are categorized under 'Nouns' and 'Mögliche Grundformen'. An advertisement for 'BOURBON LEGENDS' is visible at the top of the search results area.

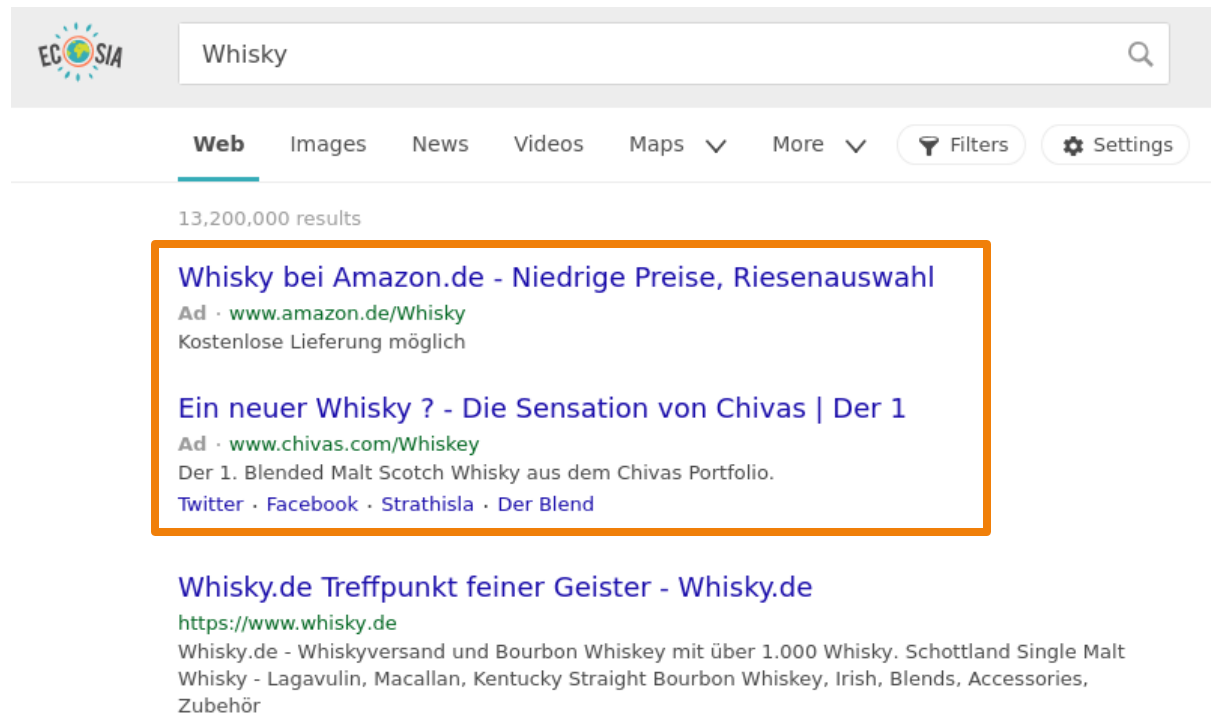
A vertical sidebar menu titled 'Dictionary Navigation' listing various language pairs and services: English ↔ German, French ↔ German, Spanish ↔ German, Italian ↔ German, Chinese ↔ German, Russian ↔ German, Portuguese ↔ German, and Polish ↔ German. Each entry includes a 'Dictionary' link and a 'Forums Trainer Courses' link.

An advertisement for Vodafone Business 400 Cable. It features a red background with white text and icons. The main offer is 'Internet & Phone Business 400 Cable' for '19€ mtl.' (monthly). A secondary offer is 'Business Booster: 50 Mbit/s Upload-Speed GRATIS!'. The Vodafone Business logo is at the bottom.

- **Programmatic Advertising: Echtzeitauktionen**
 - Webseite signalisiert: 30-35-jähriger Mann aus Leipzig mit langsamer Internetverbindung und Vorliebe für Whisky
 - Interessenten bieten automatisch auf Werbeplatz
- Deutschland: 835 Mio Euro Umsatz im Jahr 2017 [c't 2018/21, S.40]

Programmatic Advertising

- Webseite eines Händlers: Zugriff auf Kaufverhalten der eigenen Kunden
- Einsatz von Cookies:
 - Verfolgung quer durchs Internet
 - Dritte Webseite kann Werbung für Händler schalten
- Suchmaschine: Verwendung der Anfragen



The screenshot shows a search engine interface with the search term 'Whisky'. The results are categorized under 'Web' and show 13,200,000 results. Two search results are highlighted with an orange border:

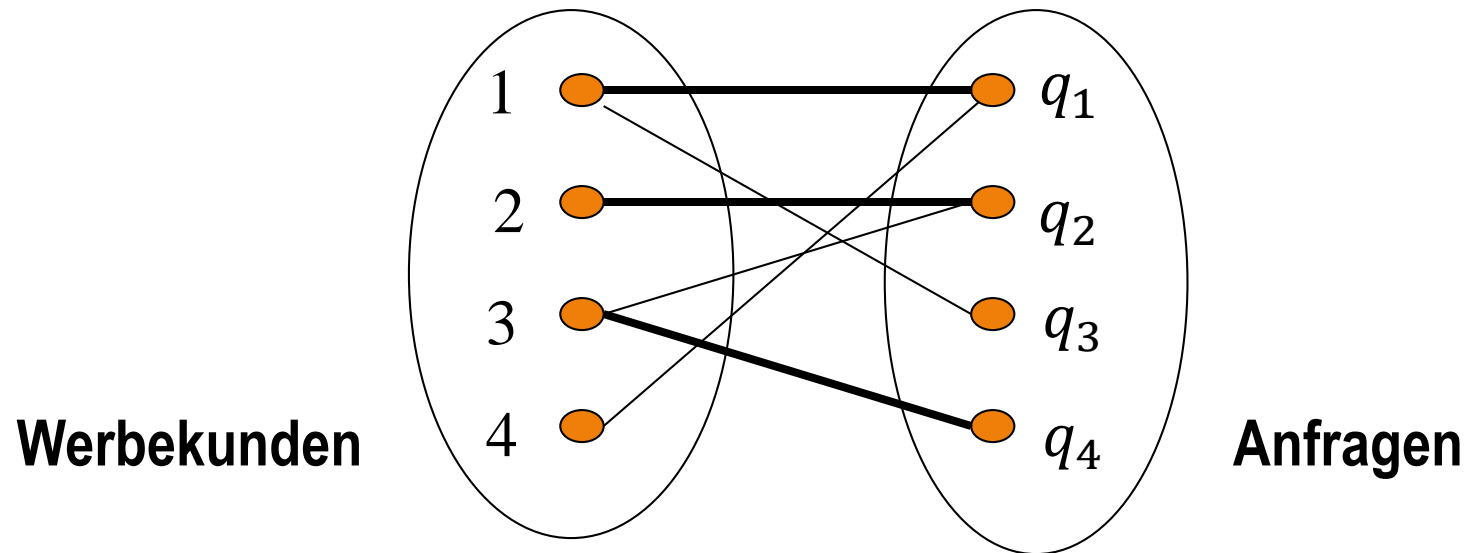
- Whisky bei Amazon.de - Niedrige Preise, Riesenauswahl**
Ad · www.amazon.de/Whisky
Kostenlose Lieferung möglich
- Ein neuer Whisky ? - Die Sensation von Chivas | Der 1**
Ad · www.chivas.com/Whiskey
Der 1. Blended Malt Scotch Whisky aus dem Chivas Portfolio.
[Twitter](#) · [Facebook](#) · [Strathisla](#) · [Der Blend](#)

Below the highlighted results, another result is visible:

- Whisky.de Treffpunkt feiner Geister - Whisky.de**
<https://www.whisky.de>
Whisky.de - Whiskyversand und Bourbon Whiskey mit über 1.000 Whisky. Schottland Single Malt Whisky - Lagavulin, Macallan, Kentucky Straight Bourbon Whiskey, Irish, Blends, Accessories, Zubehör

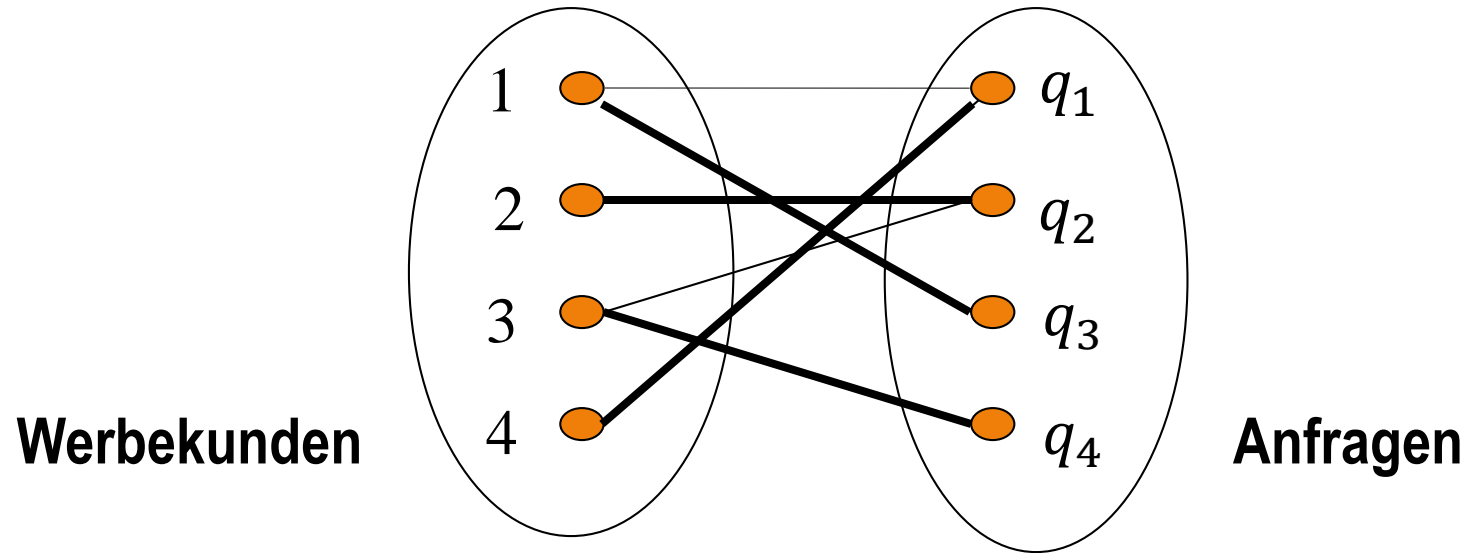
Modell

- Webseite erhält Datenstrom aus Suchanfragen $q_1, q_2, q_3 \dots$
- Mehrere Werbekunden setzen Gebot je nach Suchanfrage (Kanten)
- Webseite muss Werbekunden für Anfragen auswählen (*maximal eine Anfrage pro Werbekunde*)



Ziel: Zuordnung von Kunden zu Anfragen, so dass eine maximale Anzahl von Kunden zufrieden sind

Beste Zuordnung



Ziel: Zuordnung von Kunden zu Anfragen, so dass eine maximale Anzahl von Kunden zufrieden sind

Matching Algorithmus

- **Bipartiter Graph:** Graph aus 2 Gruppen von Knoten, wobei Kanten nur zwischen den Gruppen verlaufen
 - **Matching:** Menge von Kanten, wobei keine zwei Kanten einen gemeinsamen Knoten betreffen
 - **Maximales Matching:** eine maximale Anzahl an Kanten ist Teil des Matching
- **Ziel:** Maximales Matching für einen gegebenen bipartiten Graphen
 - Effizienter **Offline** Algorithmus (Graph vollständig bekannt): Hopcroft und Karp (https://de.wikipedia.org/wiki/Algorithmus_von_Hopcroft_und_Karp)
 - **Online:** Graph entsteht schrittweise (liegt nicht vollständig vor)
- **Online Problem:** Entscheidungen müssen augenblicklich getroffen werden, ohne die Kenntnis der zukünftigen Anfragen

Inhaltsverzeichnis

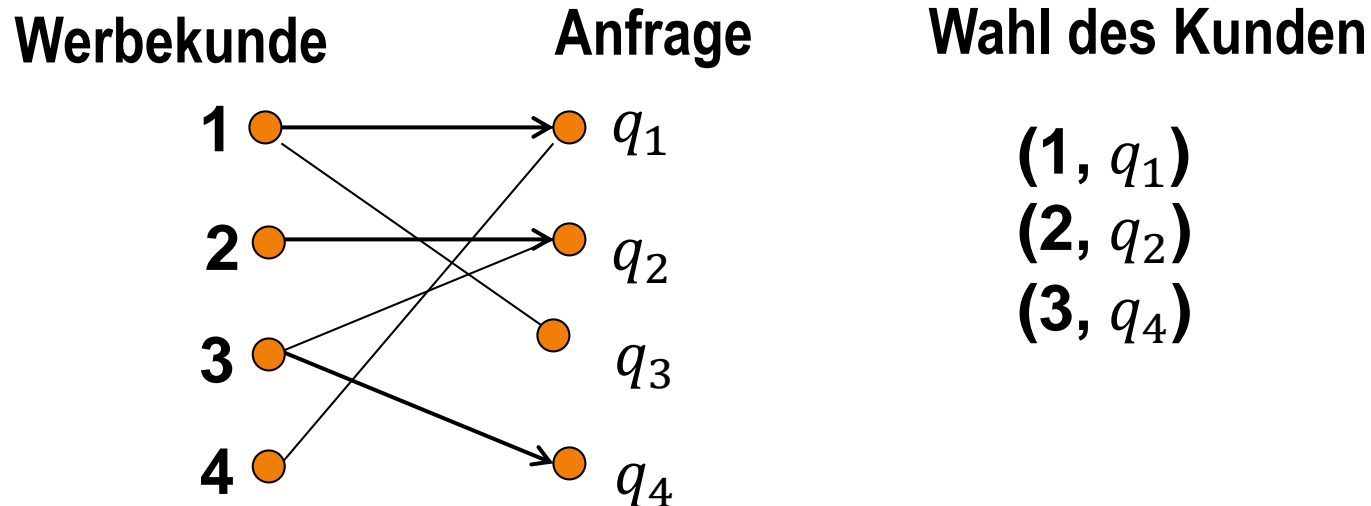
- Einführung
- Greedy Matching Algorithmus
- Balance Matching Algorithmus

Literatur: Kapitel 8 aus „Mining of Massive Datasets“: <http://www.mmds.org>

Greedy Matching

Ankommende Anfragen werden dem ersten verfügbaren Werbekunden zugeordnet

- Kunden sind geordnet
- Nimm ersten Kunden mit Gebot für Angebot



Competitive Ratio

- Wie gut ist der Greedy Algorithmus?
- Sei I eine Serie von Eingaben (z.B. Anfragen)
- Sei $M_{greedy}(I)$ das Matching, welches durch Greedy für I entsteht
- Sei $M_{opt}(I)$ ein maximales Matching für I
- Sei $|M|$ die Kardinalität von M

- **Competitive Ratio:**

$$c_{greedy} = \min_I \left(\frac{|M_{greedy}(I)|}{|M_{opt}(I)|} \right)$$

- Für jede Serie von Eingaben ist das Ergebnis von Greedy mindestens c_{greedy} mal so gut wie das optimale Ergebnis

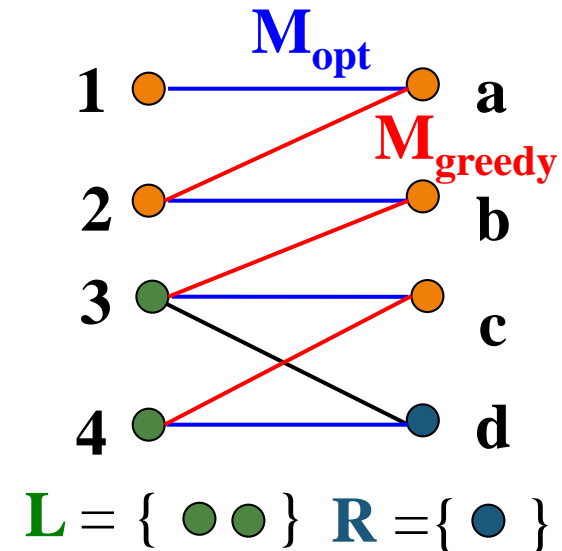
Analyse des Greedy Algorithmus

- Sei R die Menge der Knoten, die durch M_{opt} aber nicht durch M_{greedy} abgedeckt werden, d.h. $|M_{opt}(I)| \leq |M_{greedy}(I)| + |R|$
- Sei L die Menge der Knoten, die eine Kante zu Knoten aus R aufweisen und durch M_{greedy} abgedeckt werden: $|L| \leq |M_{greedy}(I)|$
- Außerdem gilt $|R| \leq |L|$
- Daraus folgt:

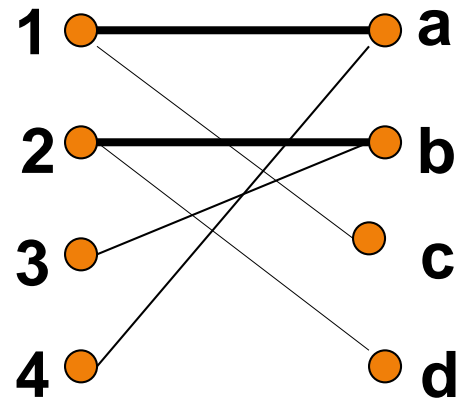
$$|M_{opt}(I)| \leq |M_{greedy}(I)| + |M_{greedy}(I)|, \text{ bzw.}$$

$$|M_{greedy}(I)| \geq \frac{1}{2} |M_{opt}(I)|$$

$$\text{d.h. } c_{greedy} = \frac{1}{2}$$



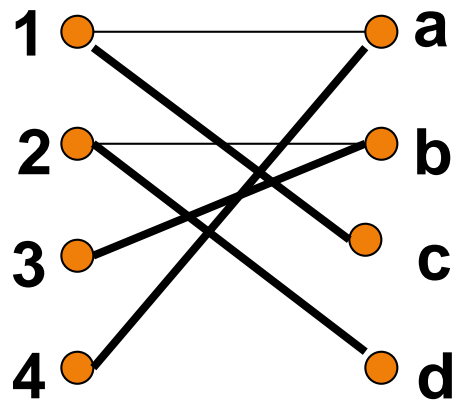
Ungünstigster Fall: Beispiel



(1,a)

(2,b)

Optimum:



Inhaltsverzeichnis

- Einführung
- Greedy Matching Algorithmus
- **Balance Matching Algorithmus**

Literatur: Kapitel 8 aus „Mining of Massive Datasets“: <http://www.mmds.org>

Werbekunden mit Budget

- Gegeben:
 1. Die Gebote von Werbekunden für Suchanfragen
 2. Konstantes *Budget* pro Werbekunde und Tag
 3. Konstanter erwarteter Gewinn pro Zuordnung
- Suche nach Menge von Werbekunden, so dass
 1. Jeder Werbekunde tatsächlich auf Suchanfrage geboten hat
 2. Jeder Werbekunde genügend Budget hat, um Klick auf Werbebanner zu bezahlen
 3. Der erwartete Gewinn maximiert wird

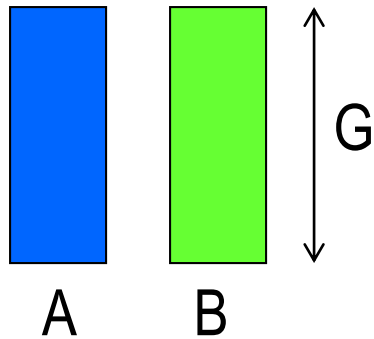
Ungünstigster Fall für Greedy

- Beispiel: Zwei Werbekunden A und B
 - A bietet auf Suchanfrage 1 und 2
 - B bietet auf Suchanfrage 1
 - Beide Werbekunden haben Budget von 4 €
 - Erwarteter Gewinn pro Zuordnung ist immer 1 €
- Reihenfolge der tatsächlichen Suchanfragen: 1 1 1 1 2 2 2 2
 - Ungünstigste Wahl durch Greedy: A A A A _ _ _ _
 - Optimal: B B B B A A A A
 - Competitive Ratio = $\frac{1}{2}$

Balance Algorithmus

- Balance Algorithmus
 - Von Mehta, Saberi, Vazirani, und Vazirani (Google Ads)
 - **Regel: Ankommende Anfragen werden dem Werbekunden mit dem derzeit größten Budget zugeordnet**
- Selbes Beispiel mit Suchanfragen: 1 1 1 1 2 2 2 2
 - Balance: A B A B A A _ _
 - Optimal: B B B B A A A A
- Allgemeiner Fall mit beliebigen Suchanfragen aber gleichem Budget G für alle Werbekunden und 1€ erwarteter Gewinn pro Anfrage:
 - Annahme: Optimale Lösung verbraucht Budgets beider Werbekunden (Gewinn: $2G$)
 - Sei x die Anzahl der Anfragen, die zwar im optimalen Fall aber nicht durch Balance zugordnet werden können
 - Erwarteter Gewinn durch Balance: $2G - x$
 - Wie groß ist x ?

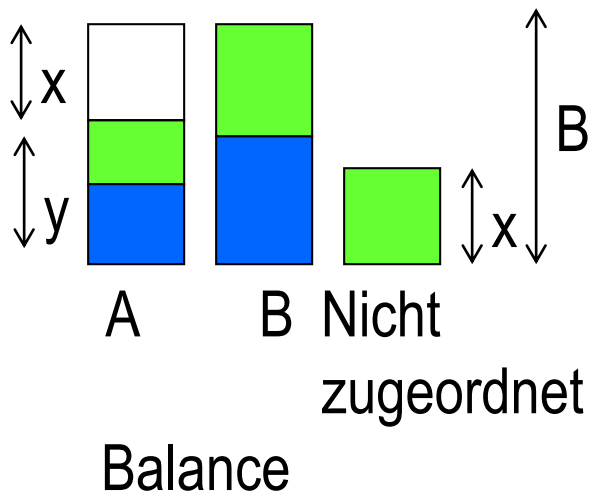
Analyse des Balance Algorithmus



- Anfragen, die im optimalen Fall dem Kunden A zugeordnet wurden
- Anfragen, die im optimalen Fall dem Kunden B zugeordnet wurden

Maximaler Gewinn: $2G$

Gewinn durch Balance: $2G - x = G + y$



Behauptung: $y \geq x$

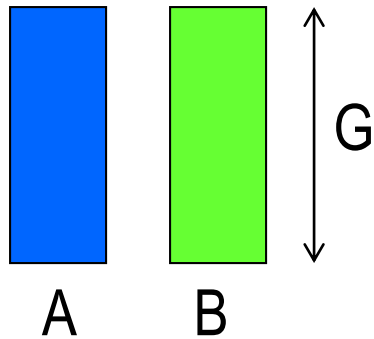
Mindestgewinn durch Balance, falls

$$x = y = \frac{G}{2}$$

Mindestgewinn: $\frac{3G}{2}$

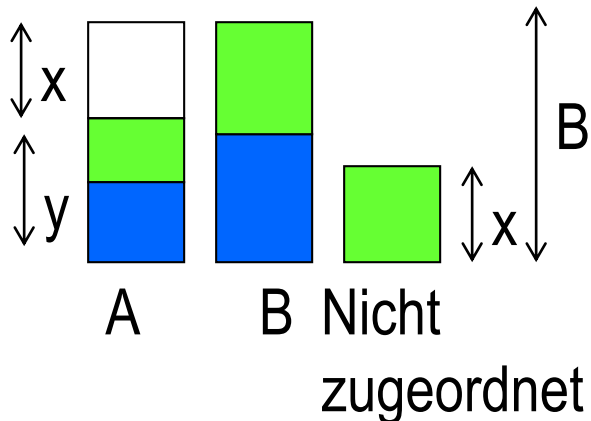
Competitive Ratio (2 Werbekunden): $\frac{3}{4}$

Analyse des Balance Algorithmus



Behauptung: $y \geq x$

- 1. Fall: Mindestens die Hälfte der blauen Anfragen werden A zugeordnet
- 2. Fall: Weniger als die Hälfte der blauen Anfragen werden A zugeordnet
 - Sei q die letzte blaue Anfrage, die B zugewiesen wurde
 - Da mehr als die Hälfte aller blauen Anfragen B zugeordnet wurden, war das Budget von B kleiner als $\frac{G}{2}$ zu diesem Zeitpunkt
 - Außerdem kann, zu diesem Zeitpunkt, das Budget von A nicht größer als das Budget von B gewesen sein, also auch kleiner als $\frac{G}{2}$
 - Daraus folgt: $y \geq \frac{G}{2}$ bzw. $y \geq x$



Balance

Balance Algorithmus

- Allgemein gilt (mehr als zwei Werbekunden):

$$c_{balance} = 1 - \frac{1}{e} \approx 0.63$$

- Es existiert kein Online Algorithmus mit höherem Competitive Ratio

- Ungünstigster Fall (mit $c_{balance} \approx 0.63$):

- Werbekunden $A_1, A_2, A_3, \dots, A_N$ mit jeweils gleichem Budget $G > N$

- Gebote:

- Suchanfrage $q_1: A_1, A_2, A_3, \dots, A_N$

- Suchanfrage $q_2: A_2, A_3, \dots, A_N$

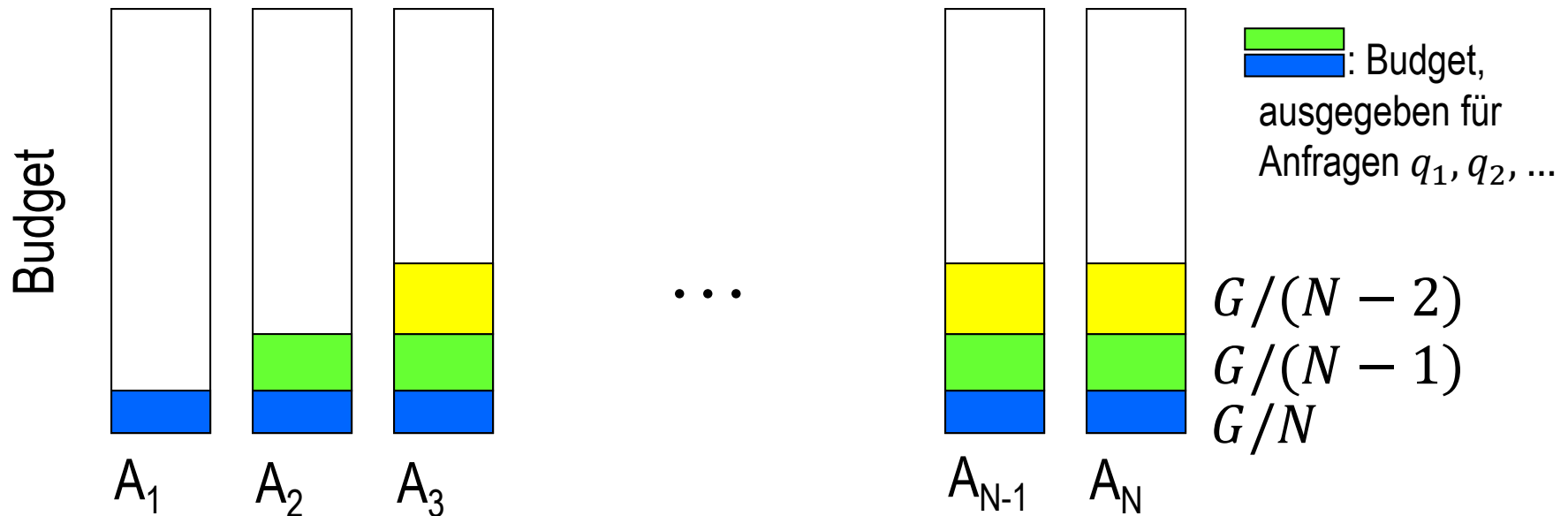
- ...

- Suchanfrage $q_N: A_N$

- Reihenfolge der Suchanfragen: $\underbrace{q_1, \dots, q_1}_{G \text{ mal}}, \underbrace{q_2, \dots, q_2}_{G \text{ mal}}, \underbrace{q_3, \dots, q_3}_{G \text{ mal}}, \dots, \underbrace{q_N, \dots, q_N}_{G \text{ mal}}$

- Optimale Lösung: $\underbrace{A_1, \dots, A_1}_{G \text{ mal}}, \underbrace{A_2, \dots, A_2}_{G \text{ mal}}, \underbrace{A_3, \dots, A_3}_{G \text{ mal}}, \dots, \underbrace{A_N, \dots, A_N}_{G \text{ mal}}$

Balance Algorithmus



Der Balance-Algorithmus verteilt die Suchanfragen gleichmäßig

- Für das verbrauchte Budget von A_k , $S_k := \sum_{i=1}^k \frac{G}{N-(i-1)}$, gilt $S_k > G$ ungefähr (Approximation nach Satz von Euler), falls $k > N \left(1 - \frac{1}{e}\right)$
- Alle Anfragen q_l mit $l > N \left(1 - \frac{1}{e}\right)$ können nicht zugeordnet werden
- Erwarteter Gewinn durch Balance: maximal $GN \left(1 - \frac{1}{e}\right)$: $c_{balance} = 1 - \frac{1}{e}$

Das Werbungsproblem

Weder Gebote noch erwarteter Gewinn ist konstant

Werbekunde	Gebot	Klickrate	Erwarteter Gewinn
A	€ 1.00	1%	1 Cent
B	€ 0.75	2%	1.5 Cent
C	€ 0.50	2.5%	1.125 Cent

Klickraten werden geschätzt aus dem vergangenen Verhalten der Nutzer

Das Werbungsproblem

- Beispiel

- Zwei Werbekunden A und B; 10 mal die gleiche Suchanfrage q

Werbekunde i	Erwarteter Gewinn pro Zuordnung x_i	Budget G_i
A	1€	110€
B	10€	100€

- Balance Algorithmus würde immer A auswählen (Erwarteter Gewinn: 10€)
- Erwarteter Gewinn bei optimaler Zuordnung zu B: 100€

- Erweiterung des Balance-Algorithmus

- Verbrauchtes Budget m_i und Anteil des verbleibenden Budgets: $f_i := 1 - \frac{m_i}{G_i}$
- **Ankommende Anfragen werden dem Werbekunden mit dem derzeit größten Wert für $x_i \cdot (1 - e^{-f_i})$ zugeordnet**
- Competitive Ratio: $1 - \frac{1}{e}$