Algorithms for Data Science: Web Advertising Algorithms

Pablo Mollá Chárlez

October 18, 2024

Contents

1 History

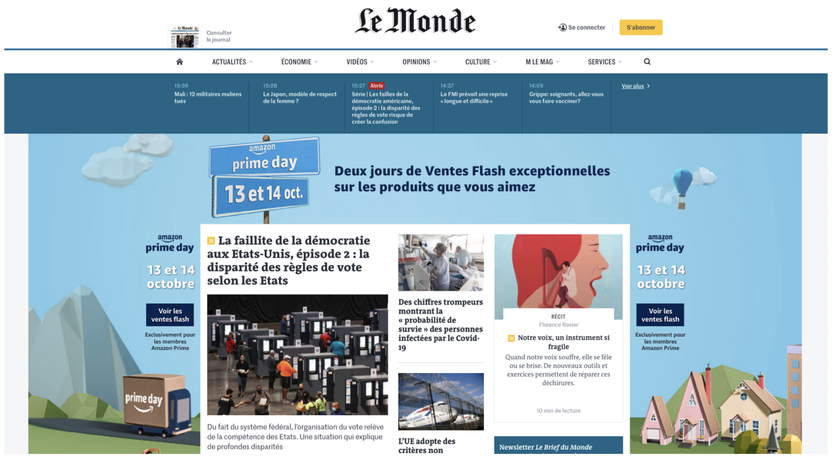
The development of online advertising has evolved significantly over the years, with distinct phases reflecting changing approaches to monetizing web content and increasing the effectiveness of ads.

1.1 First Iteration: Banner Ads (around 1995)

Initially, banner ads became popular as the main form of online advertising. Websites displayed banners, and advertisers were charged based on the number of times an ad was displayed (impressions), typically measured in units of 1,000 impressions. This model is known as CPM (Cost Per Thousand Impressions), a common approach used in traditional media like TV and print.

Early banner ads were largely untargeted, meaning they were shown to all visitors regardless of their interests or demographics. Over time, advertisers began to seek ways to target ads to specific user groups, aiming to improve relevance by tailoring content to users' age, location, or browsing behavior.

Banner ads typically had low click-through rates, meaning a small percentage of viewers actually clicked on the ad. This resulted in low return on investment (ROI) for advertisers, who were looking for more effective ways to reach potential customers.



15:59 Mali : 12 militaires maliens tués

15:28 Le Japon, modèle de respect de la femme ?

15:27 Alerte Série | Les failles de la démocratie américaine, épisode 2 : la disparité des règles de vote risque de créer la confusion

14:37 Le FMI prévoit une reprise « longue et difficile »

14:09 Grippe: soignants, allez-vous vous faire vacciner?

Voir plus

amazon prime day 13 et 14 oct.

Deux jours de Ventes Flash exceptionnelles sur les produits que vous aimez

amazon prime day 13 et 14 octobre Voir les ventes flash Exclusivement pour les membres Amazon Prime

La faillite de la démocratie aux Etats-Unis, épisode 2 : la disparité des règles de vote selon les Etats

Du fait du système fédéral, l'organisation du vote relève de la compétence des Etats. Une situation qui explique de profondes disparités

Des chiffres trompeurs montrant la « probabilité de survie » des personnes infectées par le Covid-19

L'UE adopte des critères non

RÉCIT
Florence Rosier
Notre voix, un instrument si fragile
Quand notre voix souffre, elle se fêle ou se brise. De nouveaux outils et exercices permettent de réparer ces déchirures.
10 min de lecture

Newsletter *Le Brief du Monde*

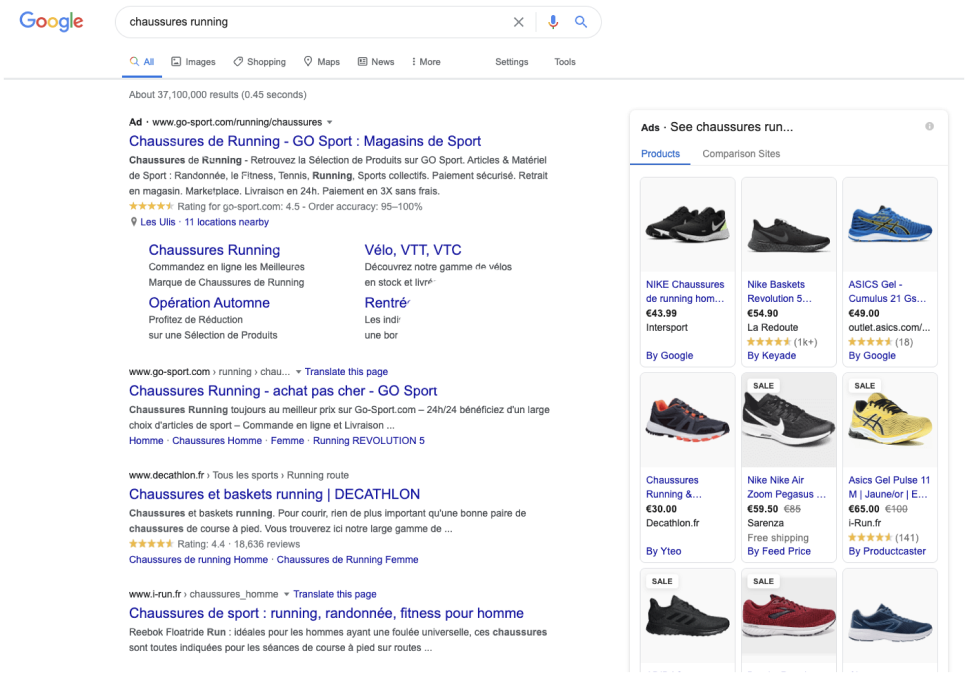
amazon prime day 13 et 14 octobre Voir les ventes flash Exclusivement pour les membres Amazon Prime

1.2 Second Iteration: Ads on Search Results (around 2001)

As search engines became central to internet use, a new model emerged for advertising—ads linked to search results. Advertisers would bid on keywords, meaning they would pay to have their ad displayed when users searched for specific terms.

Unlike the CPM model, this approach charged advertisers only if a user clicked on the ad, known as a pay-per-click (PPC) model. This significantly improved the ROI since advertisers paid only for actual engagement.

Google adopted this advertising model around 2002, launching its AdWords platform. This system allowed advertisers to create ad campaigns targeted at specific keywords, enabling them to appear in sponsored search results.



About 37,100,000 results (0.45 seconds)

Ad · www.go-sport.com/running/chaussures

Chaussures de Running - GO Sport : Magasins de Sport

Chaussures de **Running** - Retrouvez la Sélection de Produits sur GO Sport. Articles & Matériel de Sport : Randonnée, le Fitness, Tennis, **Running**, Sports collectifs. Paiement sécurisé. Retrait en magasin. Marketplace. Livraison en 24h. Paiement en 3X sans frais.

★★★★★ Rating for go-sport.com: 4.5 - Order accuracy: 95–100%

Les Ulis · 11 locations nearby

Chaussures Running
Commandez en ligne les Meilleures Marque de Chaussures de Running

Vélo, VTT, VTC
Découvrez notre gamme de vélos en stock et livré

Opération Automne
Profitez de Réduction sur une Sélection de Produits

Rentré
Les indi
une bor

www.go-sport.com › running › chau... Translate this page

Chaussures Running - achat pas cher - GO Sport

Chaussures Running toujours au meilleur prix sur Go-Sport.com – 24h/24 bénéficiez d'un large choix d'articles de sport – Commande en ligne et Livraison ...

Homme · Chaussures Homme · Femme · Running REVOLUTION 5

www.decathlon.fr › Tous les sports › Running route

Chaussures et baskets running | DECATHLON

Chaussures et baskets **running**. Pour courir, rien de plus important qu'une bonne paire de **chaussures** de course à pied. Vous trouverez ici notre large gamme de ...

★★★★★ Rating: 4.4 · 18,636 reviews

Chaussures de running Homme · Chaussures de Running Femme

www.i-run.fr › chaussures_homme Translate this page

Chaussures de sport : running, randonnée, fitness pour homme

Reebok Floatride **Run** : idéales pour les hommes ayant une foulée universelle, ces **chaussures** sont toutes indiquées pour les séances de course à pied sur routes ...

Ads · See chaussures run...

Products | Comparison Sites

NIKE Chaussures de running hom...
€43.99
Intersport
By Google

Nike Baskets Revolution 5...
€54.90
La Redoute
★★★★★ (1k+)
By Keyade

ASICS Gel - Cumulus 21 Gs...
€49.00
outlet.asics.com/...
★★★★★ (18)
By Google

Chaussures Running &...
€30.00
Decathlon.fr
By Yteo

SALE
Nike Nike Air Zoom Pegasus ...
€59.50 ~~€85~~
Sarenza
Free shipping
By Feed Price

SALE
Asics Gel Pulse 11 M | Jaune/or | E...
€65.00 ~~€100~~
i-Run.fr
★★★★★ (141)
By Productcaster

SALE

SALE

1.3 Web 2.0 and Performance-Based Advertising

As part of the Web 2.0 era, performance-based advertising became a multi-billion-dollar industry. It enabled companies to better allocate their advertising budgets by targeting users who were more likely to be interested in their products or services.

One major problem faced by search engines was deciding which ads to show for a given search query. Another challenge for advertisers was determining the right keywords to bid on and how much to bid to ensure that their ads appeared at the top of search results. These issues are tied to concepts from computational game theory, where strategies are devised for optimal bidding in a competitive environment.

The shift from banner ads to search-based advertising marked a major turning point in digital marketing, leading to more sophisticated targeting and pricing models that continue to shape online advertising today.

2 The Online Matching Problem

The concepts of greedy algorithms and competitive ratios play a role in making real-time decisions when facing constraints and uncertainties in performance-based advertising.

An online optimization problem is a type of optimization problem where decisions must be made sequentially, based only on the information available at that time, without knowing future inputs. The goal is to either maximize or minimize an objective function as new data arrives. Unlike **offline optimization**, where all the data is available beforehand, online optimization deals with uncertainty and dynamic changes.

We will study 2 main approaches to solve the online matching problem, the Greedy Algorithm and the Balance Algorithm.

2.1 Matching Problem

The matching problem involves finding a way to connect pairs of elements in a set, ensuring certain conditions are met. Let's first define a concept to understand and visualize the matching problem.

2.1.1 Bipartite Graph

A $\boxed{\text{bipartite graph}}$ is a graph $G(V_1 \cup V_2, E)$, where:

- There are two disjoint sets of nodes, V_1 and V_2.
- Edges only connect nodes from V_1 to nodes in V_2.

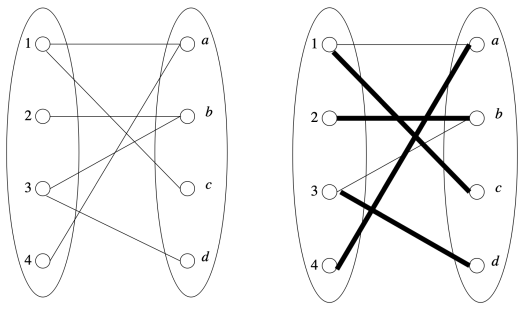
For instance, consider a bipartite graph with nodes $V_1 = \{1, 2, 3, 4\}$ and $V_2 = \{a, b, c, d\}$ and edges defined as E $=\{(1, a), (1, c), (2, b), (3, b), (3, d), (4, a)\}$.

2.1.2 Matching in a Bipartite Graph

A matching is a subset of edges such that no node is connected by more than one edge in the matching. There are different types:

- $\boxed{\text{Perfect Matching}}$ Every node is included in the matching.
- $\boxed{\text{Maximal Matching}}$ The largest possible matching, with the most number of edges.

For the previous example graph, a maximal matching is: $\{(1, c), (2, b), (3, d), (4, a)\}$.



In the **offline case**, algorithms can use all the information about the graph to find a maximal matching. Efficient algorithms for finding maximal matchings can run in $O(n^2)$, where $n = |E|$.

For the **online case**, edges are presented one by one, and decisions must be made in the order they arrive.

2.2 Greedy Algorithm for Online Optimization

The greedy algorithm is a common approach to online optimization problems. In this method, decisions are made locally by optimizing based on the current data and previous decisions, without considering future possibilities. It aims for a quick and locally optimal choice at each step.

- $\boxed{\textbf{Not Always Optimal}}$: While a greedy algorithm can produce good results, it doesn't always yield the best solution when compared to offline algorithms, which have access to the full dataset.
- $\boxed{\textbf{Competitive Ratio}}$: The performance of a greedy algorithm in an online setting is measured using the competitive ratio, defined as:

$$c = min\{\frac{|M_g|}{|M_o|}\}$$

  Here, $|M_g|$ is the objective function value obtained by the greedy algorithm, and $|M_o|$ is the optimal value obtained by an offline algorithm. The competitive ratio gives a lower bound on how well the online solution compares to the best possible offline solution.

2.2.1 Competitive Ratio of Greedy Algorithm

The competitive ratio analysis provides insight into how the performance of the greedy algorithm compares to the optimal matching. Let M_o be the maximal matching, M_g be the greedy matching, L represents the left nodes that are matched in M_o but not in M_g and R the right nodes connected to any node in L.

2.2.1.1 Claim 1

$\boxed{\textbf{Claim}}$ Every node in R is matched in M_g.

$\boxed{\textbf{Proof}}$ This can be shown by contradiction. Suppose a node in R is not matched in M_g, there must exist an edge connecting it to a node in L, which would imply it could be matched, contradicting the assumption.

2.2.1.2 Claim 2

$\boxed{\textbf{Claim}}$ The competitive ratio of the greedy algorithm is $\frac{1}{2}$.

$\boxed{\textbf{Proof}}$ We derive the competitive ratio by establishing relationships between these sets to understand how close the greedy algorithm's solution can get to the optimal solution.

1. $\boxed{|M_o| \leq |M_g| + |L|}$

   The optimal matching M_o can only include the edges in M_g plus additional edges connected to nodes in L, which are unmatched by the greedy algorithm. This inequality shows that the optimal matching can be decomposed into edges that appear in the greedy matching and those connected to unmatched nodes in L.

2. $\boxed{|L| \leq |R|}$

   In the optimal matching M_o, each node in L is matched with a distinct node in R, meaning there are at least as many nodes in R as there are in L.

3. $\boxed{|R| \leq |M_g|}$

   Every node in R is matched in the greedy matching M_g, implying that the number of nodes in R cannot exceed the size of the greedy matching.

4. We combine the inequalities to derive the lower bound on the competitive ratio:

   - Start with $|M_o| \leq |M_g| + |L|$.
   - Substitute $|L| \leq |R|$, giving $|M_o| \leq |M_g| + |R|$.
   - Substitute $|R| \leq |M_g|$, resulting in $|M_o| \leq |M_g| + |M_g| = 2|M_g|$.

   Thus, $|M_g| \geq |M_o|/2$, which indicates that the greedy matching is at least half the size of the optimal matching. This establishes that the competitive ratio is at least 1/2.

   While 1/2 is a lower bound, there are cases where the greedy algorithm's matching is exactly half the size of the optimal matching. These cases serve as counterexamples, demonstrating that the greedy algorithm cannot always do better than achieving half the size of the optimal solution. Therefore, 1/2 is also the upper bound for the competitive ratio. This means that, in the worst-case scenario, the greedy algorithm's performance will be no better than half of what the optimal offline algorithm could achieve.

2.2.1.3 Example: Competitive Ratio

Let's go through an example to clarify the roles of M_o, M_g, L, and R in the context of the online matching problem. Let's reuse the bipartite graph with two sets of nodes and the edges described in the previous pictures.

- Left set $V_1 = \{1, 2, 3, 4\}$
- Right set $V_2 = \{\text{a, b, c, d}\}$
- Edges connecting V_1 and V_2: {(1, a), (1, c), (2, b), (3, b), (3, d), (4, a)}

1. **Step 1: Greedy Matching (M_g)**

   Suppose the edges arrive in the order given above, and we use a greedy algorithm to build the matching by adding an edge only if neither endpoint is already matched.

   The edges are processed in this order:

   (a) (1, a): Add to M_g because both 1 and a are unmatched.
   (b) (1, c): Skip because node 1 is already matched.
   (c) (2, b): Add to M_g because both 2 and b are unmatched.
   (d) (3, b): Skip because node b is already matched.
   (e) (3, d): Add to M_g because both 3 and d are unmatched.
   (f) (4, a): Skip because node a is already matched.

   The resulting greedy matching is: $M_g = \{(1, a), (2, b), (3, d)\}$

2. **Step 2: Optimal Matching (M_o)**

   Now, let's find the optimal offline matching (the maximum matching possible with full knowledge of the graph). One possible optimal matching is: $M_o = \{(1, c), (2, b), (3, d), (4, a)\}$

3. **Step 3: Identifying L and R**

   Now, we need to determine the sets L and R:

   - L: Nodes in the left set V_1 that are matched in M_o but not in M_g. In this case, node 4 is matched in M_o (to a), but not in M_g. So, $L = \{4\}$.
   - R: Nodes in the right set V_2 that are connected to any node in L via an edge in the original graph. Node 4 is connected to node a. Therefore, $R = \{a\}$.

4. $\boxed{\textbf{Step 4: Competitive Ratio}}$

   - $|M_o| = 4$: The optimal matching has four edges.
   - $|M_g| = 3$: The greedy matching has three edges.
   - $|L| = 1$: There is one unmatched node in L.
   - $|R| = 1$: There is one corresponding node in R.

   The inequalities derived earlier can be verified:

   - $|M_o| \leq |M_g| + |L| \implies 4 \leq 3 + 1$, which holds true.
   - $|L| \leq |R| \implies 1 \leq 1$, which is true.
   - $|R| \leq |M_g| \implies 1 \leq 3$, which is also true.

   Finally, $|M_g| \geq |M_o|/2 \implies 3 \geq 4/2 = 2$, confirming that the competitive ratio of the greedy algorithm is indeed at least $1/2$.

In this example, the greedy solution captured 3 out of 4 possible matches, illustrating the concept of competitive ratio in the context of online matching problems.

2.3 Balance Algorithm

The online matching problem involves assigning items (like queries or tasks) to available resources (such as advertisers or servers) as they arrive, without knowing future items in advance. We previously saw how a greedy algorithm can perform this assignment in a straightforward manner by matching each incoming item to an available resource, resulting in a competitive ratio of $\frac{1}{2}$.

The BALANCE algorithm improves upon the greedy approach by considering the remaining budget of each advertiser. Let's see how it works:

1. Assign the query to the advertiser with the most budget left. The intuition is that advertisers with higher budgets remaining are less likely to run out of budget soon, ensuring that the budget is spread across more queries.

2. Competitive ratio is $\frac{3}{4}$. This means that the revenue obtained by BALANCE is at least 75% of the optimal offline solution's revenue.

3. Deterministic tie-breaking. If two advertisers have the same remaining budget, a deterministic method must be used to decide which advertiser gets the query to ensure consistent performance.

2.3.1 Competitive Ratio Derivation for BALANCE: 2 Bidders Case

Let's explore the competitive ratio for the BALANCE algorithm in the case of two advertisers, A_1 and A_2, with budgets B.

1. Assume BALANCE must exhaust the budget of at least one bidder. Suppose A_2's budget is exhausted first, while A_1 still has some budget left.

2. Divide queries into cases:

   - If at least half of the queries are assigned to A_1, then $y \geq B/2$.
   - If more than half of the queries are assigned to A_2, the remaining budget for A_2 is less than $B/2$, making $x \leq B/2$.

3. Result: At the minimal case of $x = y = B/2$, the revenue from BALANCE is $3B/2$, while the optimal revenue is $2B$, resulting in a competitive ratio of $3/4$.

2.3.2 General Case for Multiple Bidders

When there are multiple advertisers (N bidders), the competitive ratio for BALANCE in the general case is $(1 - \frac{1}{e}) \approx 0.63$. Here's how it works:

1. $\boxed{\textbf{Worst-case scenario setup}}$ - Advertisers $A_1, A_2, \ldots, A_N$ each have a budget B. - Queries arrive in multiple rounds, with the number of queries in each round proportional to N.

2. $\boxed{\textbf{Revenue Bound Analysis}}$ - In each round, the advertisers with higher budgets left receive a higher proportion of the queries. - As the rounds progress, some advertisers exhaust their budgets earlier, limiting BALANCE's ability to optimally allocate remaining queries.

3. $\boxed{\textbf{Competitive Ratio Calculation}}$ - The smallest k at which BALANCE stops generating revenue due to budget constraints occurs around $N(1 - \frac{1}{e})$, providing a total revenue bound of $B \times N(1 - \frac{1}{e})$.

2.3.3 Generalized BALANCE Algorithm

To address cases with arbitrary bids:

1. Define the fraction of the leftover budget: $f_i = 1 - \frac{m_i}{b_i}$, where m_i is the amount spent so far, and b_i is the original budget.

2. Use a scoring function $\psi_i(q) = x_i(1 - e^{-f_i})$, where x_i is the bid for query q.

3. Allocate the query to the bidder with the highest $\psi_i(q)$.

The generalized BALANCE algorithm still achieves a competitive ratio of $1 - \frac{1}{e}$ and is the one implemented on the Python Notebook.