SIMILARITY AND NOVELTY METRICS: A MACHINE LEARNING FRAMEWORK FOR AUDIENCE EXTENSION

Sarthak Pattnaik and Eugene Pinsky

Department of Computer Science, Metropolitan College, Boston University, Boston

ABSTRACT

In the domain of digital advertising, a principal imperative is the precise identification and engagement of a target audience—comprising both extant consumers identified from historical data and potential prospects convertible into future patrons. A persistent and substantive challenge in this endeavor lies in constructing targeting constraints that not only capture existing behavioral patterns but also extrapolate toward high-propensity yet unobserved audience segments. This strategic expansion, commonly designated as audience extension, has conventionally been addressed through greedy cover algorithms, which prioritize audience volume to the exclusion of nuanced performance indicators. In this study, we present a methodological augmentation of the greedy framework by incorporating dual performance metrics—similarity and novelty—as evaluative criteria. The proposed algorithm introduces a multi-objective optimization framework that facilitates the judicious expansion of audience segments while preserving representational fidelity to the original cohort. We empirically substantiate the efficacy of our framework through multiple case studies, demonstrating its superiority in balancing quantitative performance with qualitative audience alignment.

KEYWORDS

Digital advertising, audience extension, greedy algorithm, novelty metric, targeted marketing

1. INTRODUCTION

Over the years, the landscape of advertising techniques has undergone a significant transformation. Traditional advertising approaches, which relied on broadcasting campaigns to large audiences, have lost their effectiveness in the face of changing consumer behavior, technology advances ([1], [2], [3]) and the use of machine learning techniques ([4], [5], [6], [7]). In response to this shift, digital advertising techniques have taken center stage, with a major portion of advertising budgets now allocated to precisely target audiences at the right time and place [8]. Data management platforms (DMPs) have emerged as indispensable tools in this new era of advertising, enabling the collection and integration of vast amounts of data from various sources [9]. DMPs store valuable information such as historical campaign data, demographics, credit scores, and more, empowering advertisers with insights to make informed decisions about their audience targeting strategies [10]. A simple illustration in Figure 1 shows a sample advertiser segment.

DOI : 10.5121/ieij.2025.9201

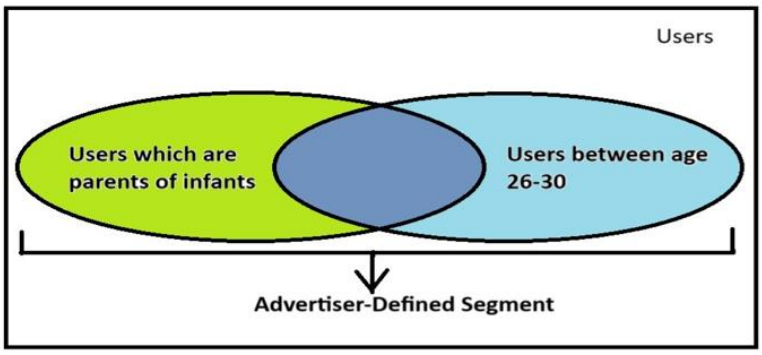


Figure 1: An example of an advertiser segment

In this context, the problem of audience extension has gained prominence ([11], [12], [13]). Advertisers often encounter situations where manually created audience segments cover only a fraction of the total target population [14]. As a result, they need to expand their audience reach to meet desired performance metrics [15]. Audience extension refers to creating a new segment that resembles the initial audience cover but also includes novel members with the potential to become future customers [16]. This problem can be viewed as a general machine learning problem where we look for patterns or classification based on similarity and difference metrics ([17], [18]).

Among the common approaches to audience extension is the ”greedy cover” algorithm, which is widely used in digital advertising ([19], [20]). However, this algorithm has limitations, and researchers have identified several pitfalls that can hinder its performance ([21], [22]). Therefore, it is essential to investigate the shortcomings of the existing algorithm and explore modifications that can lead to optimal results in addressing the audience extension problem [23]. This research aims to examine the ”greedy cover” algorithm and its effectiveness for audience extension. By reviewing existing literature and conducting empirical analysis, this study seeks to provide insights into the strengths and weaknesses of the algorithm, along with potential enhancements to achieve superior audience targeting outcomes ([24], [25]).

This paper is organized as follows: In Section 2, we outline the existing literature that is available on this domain. In section 3 we explain the mathematical concepts that play a crucial role in the formulation of the proposed approach. Section 4, we define our proposed methodology for audience extension. In Section 5, we provide three examples to illustrate and compare the performance of the greedy cover algorithm and the similarity-novelty-based approach for audience extension. Finally, in Section 6, we summarize the pertinent findings from our research that generalize the results that we have obtained.

2. RELATED WORK

2.1. What is Audience Extension?

Consider the global e-commerce giant Amazon and its strategic foray into a new market. When Amazon expands to a new region, it faces the challenge of building a customer base from scratch while ensuring a high engagement rate and conversions. Traditional marketing methods, although reliable to some extent, do not provide the granularity needed for targeted outreach in a new market.

Here is how Amazon employs advanced audience extension techniques to overcome these challenges, ensuring a successful market entry. Imagine Amazon launching its services in a country

without a prior customer base. Amazon creates an initial audience segment based on basic demographics, online behavior, and product interests to maximize its reach. However, this segment is relatively small, considering the vast population of the new market. To thrive, Amazon needs to expand this segment intelligently, targeting individuals who are likely to buy and become long-term, loyal customers. Amazon utilizes its wealth of data and machine learning algorithms to analyze many factors. These include social media interactions, search queries, wish lists, and even cursor movements on their platforms. By processing this data, Amazon identifies patterns that indicate potential customers' preferences, interests, and purchasing intent. Machine learning algorithms predict behaviors, allowing Amazon to pinpoint individuals likely to engage and convert [1]. With the insights gained, Amazon tailors its marketing strategies. Personalized product recommendations are sent to the identified potential customers. These recommendations are not generic; they are meticulously curated based on individual preferences, ensuring relevance and resonance. This tailored approach significantly increases the likelihood of click-through and conversions, as customers feel a personal connection to the offerings [2]. Amazon doesn't stop at the initial outreach. Through real-time monitoring of customer interactions, Amazon continuously adapts its strategies. For instance, if a particular audience segment shows more interest in electronics, Amazon fine-tunes its outreach efforts to emphasize technology products. This adaptability ensures that marketing efforts remain relevant and compelling, maximizing the return on investment (ROI) [8].

Amazon employs a robust analytics system to measure the success of its audience extension efforts. Metrics such as click-through rates, conversion rates, and customer lifetime value are meticulously tracked. Amazon compares these metrics with those from markets where similar strategies were not employed. By quantifying the impact, Amazon validates the effectiveness of its audience extension techniques empirically [19]. Beyond immediate sales, Amazon's strategic audience extension efforts have a profound long-term impact. Satisfied customers from the initial outreach phase tend to become loyal patrons. They make repeat purchases and contribute positively to Amazon's brand reputation through word-of-mouth and online reviews. Amazon's focus on quality audience extension thus builds a robust customer ecosystem, ensuring sustained growth and market dominance in the new region.

The Amazon example underscores the power of advanced audience extension in reshaping mar- ket entry strategies. By harnessing big data, machine learning, and personalized experiences, Amazon transforms a mere audience segment into a thriving customer base. This case study demonstrates that in the digital age, precision and personalization in audience extension are not just advantages; they are imperatives for businesses aiming to expand and thrive in new markets. Amazon's success story serves as a testament to the transformative potential of applying machine learning to strategic audience extension efforts in global e-commerce.

2.2. Greedy Cover Algorithm

We start with the description of the greedy cover algorithm presented in [26]. For advertisers, it is crucial to explicitly define a range of quantitative metrics by which the extended audience segment should be evaluated. In the majority of instances, advertisers seek an expanded audience size that surpasses a specific threshold while also achieving elevated retention rates compared to the initial coverage, all accompanied by notably improved performance metrics such as clicks and conversions ([21], [27]).

$$\mathrm{sim}(S, S') > \alpha, \qquad \mathrm{perf}(S') - \mathrm{perf}(S) > \beta, \qquad \text{and} \qquad |\mathrm{aud}(S \cup S')| \gg |\mathrm{aud}(S)|$$

The predominant approach frequently employed for audience extension is the greedy cover algorithm. This method is an extension of the greedy set cover algorithm presented in this. Before delving into the workings of this algorithm, it is imperative to explain a few core concepts under-

pinning audience segments. Let us assume that the advertiser has delineated an audience segment based on a set of features $\{c_1, c_2, c_3, \ldots, c_n\}$. Consequently, any potential audience member belonging to this segment must encompass these features. However, this does not necessarily dictate that these features constitute the entirety of the audience members' attributes. The greedy cover algorithm aspires to encompass the initial audience coverage while augmenting the audience's scale to reach a predetermined threshold. The procedure employed in the greedy approach is explicated in Algorithm 1.

Algorithm 1 Greedy Cover algorithm to extend audience

Require: Initial Segment S, Inventory of available segments Ω, required audience size N
Ensure: Extended audience cover A_{ext}
1: Sort available segments in Ω in descending order by size
2: Initialize $A_{ext} = \{\}$
3: **for** each segment $S' \in \Omega$ with $S' \supseteq S$ **do**
4: **if** $|S'| \geq N$ **then** Add S' to A_{ext}
5: **end if**
6: **end for**
return A_{ext}

Although the greedy cover algorithm is a substantially easy-to-apply algorithm, it may not provide the advertiser with the best audience cover. Here are some of the pitfalls of the greedy cover algorithm:

- The algorithm does not consider the percentage of novel audience added the advertising segment. The algorithm accepts the extended audience cover as long as it covers the initial segment and surpasses the cumulative audience size required by the advertiser.
- The algorithm fails to consider the cost for a certain number of impressions when incorporating an extended audience cover.
- The algorithm does not compare the initial audience and the extended audience cover with respect to various performance metrics (clicks, return on investments, conversions, etc.)

Therefore, the subsequent sections of this paper endeavor to devise an approach that would perform the imperative task of overcoming the limitations of the greedy cover algorithm. To perform a quantitative analysis of the novel audience added, we use various statistical measures, which are described in detail in the next section [19].

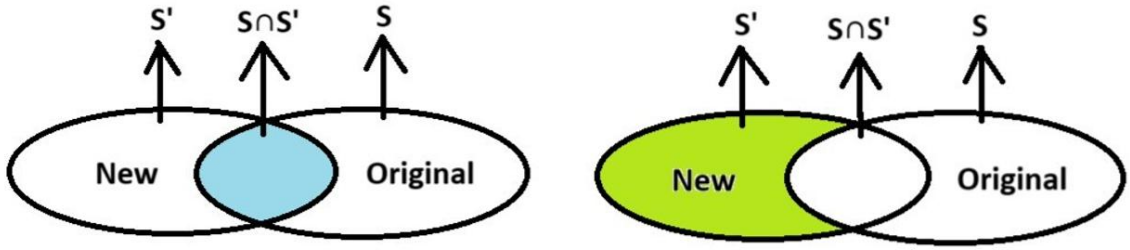
3. SIMILARITY AND NOVELTY METRICS

When extending an audience segment, two significant quantitative variables come into play, determining the viability of the extended segment for achieving better performance compared to the initial coverage ([19], [28]). The term *similarity* denotes a quantitative estimate indicating the degree to which the original audience coverage is preserved within the extended segment [27]. On the other hand, *novelty* represents an estimation of new audience members included in the extended segment who were not part of the initial coverage. To express these concepts mathematically, we denote the original coverage and the extended audience segment as S and S', respectively. The mathematical formulations for the similarity and novelty metrics are as follows:

$$\text{Sim}(S, S') = \frac{|S \cap S'|}{|S|} \quad \text{and} \quad \text{Nov}(S, S') = 1 - \frac{|S \cap S'|}{|S'|}$$

In a previous research paper [29], a similar methodology was employed where the similarity between the

advertisement campaign and multiple campaigns, which are part of the historical dataset, were calculated using the Hamming and the Jaccards distance. A subset of nearest neighbors is formed, which consists of historical campaigns that have a value of similarity below a certain threshold. In this way, the algorithm augments the cumulative impressions of the target advertisement campaign to meet the demands of the advertiser. The approach that we propose, however, is aimed at providing the best-extended cover that covers the original customers while at the same time adding candidates who have the potential to become customers in the future. Ideally, advertisers seek a balanced relationship between these two metrics. The extended audience segment should encompass a substantial portion of the initial coverage while simultaneously introducing novel audience members [21]. Figure 2 illustrates the similarity and novelty between the original segment (S) and the new audience segment (S').



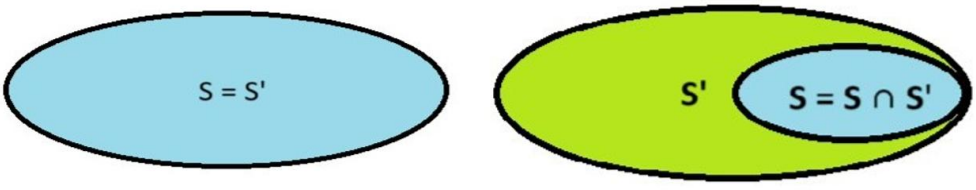
(a) Similarity (b) Novelty

Figure 2: Similarity between original segment S and novel segment S'

The performance of the extended audience segment is predicated on the similarity and novelty value. Advertisers determine whether to accept an extended audience coverage based on these performance metrics. The value of both these statistical metrics lies between 0 and 1. When the value of the similarity metric is 1 and the novelty metric is 0, the extended audience cover overlaps with the initial cover. However, when the novelty metric is 1 and the similarity metric is 0, then the advertiser would target a completely new audience base without retaining any existing customers. Both these extreme situations are worst-case scenarios for advertisers, as in the first case, the result is more or less similar to the results obtained through the greedy cover algorithm. On the other hand, in the second case, the advertiser would be losing its customer base to a new set of audiences that may or may not provide the desired results. Apart from these far-end conditions, there are a multitude of cases that may come to the foreground. These are described below in detail:

- **Case I: Maximum Similarity:**

There are two subcases, illustrated in Figure 3



(a) No Novelty (b) Novelty

Figure 3: High Novelty

1. no novelty: ($S = S'$) This is illustrated in subfigure (a). The original and the extended (new)

segment are exactly the same. The similarity metric is 1 while the novelty metric is 0.

2. maximum novelty ($S = S \cap S'$). This is illustrated in subfigure (b) of Figure 3. The extended segment covers the original segment, along with adding a significant number of novel audience members. The similarity metric is 1 and the novelty metric is quite high.

- **Case II: Low Similarity, No Novelty ($S' \subset S$):**

This case is illustrated in Figure 4

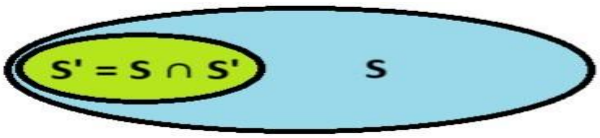


Figure 4: Low Similarity, No Novelty

The extended segment is a subset of the original segment, therefore there is abysmal similarity and no novelty. The similarity metric is low and the novelty metric is 0.

- **Case III: No Similarity, Maximum Novelty ($S \cap S' = \emptyset$)**

This is illustrated in Figure 5

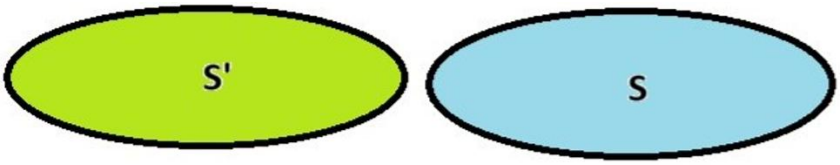
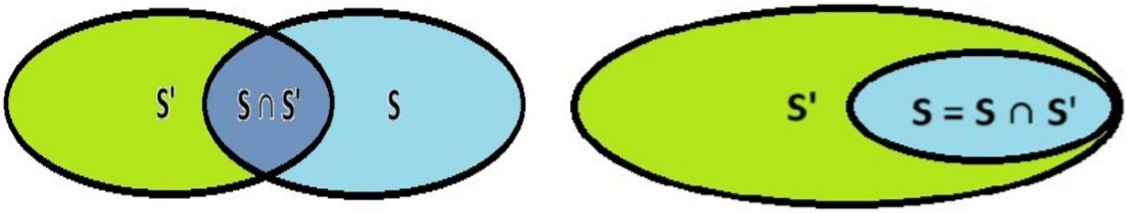


Figure 5: No Similarity, Max Novelty

The initial cover and the extended audience segment are disjoint. Here, the value of the similarity metric is 0 but the novelty metric is the 1.

- **Case IV: Some Similarity, Some Novelty (($S \cap S' \neq \emptyset$) and ($S \cap S' \neq S$) and ($S \cap S' \neq S'$))**

When there is some overlap between the initial cover and the extended segment, the values of the similarity and the novelty metrics exist between the two extremities. As one increases, the other decreases. This is illustrated in Figure 6



(a) Low Similarity, High Novelty (b) High Similarity, Low Novelty

Figure 6: Some Similarity and Some Novelty

Therefore, there are two broad sub-cases that exist when neither of the two metrics are leaning towards extreme values. The detailed description of the sub-cases are mentioned below:

1. Low Similarity, High Novelty: This is shown in the left subfigure (a) of Figure 6. The extended segment has a small overlap with the original segment. Therefore, the novelty metric is high with a low value of the similarity metric.
2. High Similarity, Low Novelty: This is shown in the right subfigure (b) of Figure 6. The original segment overlaps significantly with the new segment. This is exactly the opposite of case IV, here the similarity metric is high whereas the novelty metric is low.

After analyzing all five scenarios, we can surmise that ideally, advertisers would want the extended audience segment to cover the initial audience at the same time, add a significant amount of novel audience members. Therefore, the advertiser wants the extended segment to achieve maximum similarity (similarity metric should be ideally 1) along with a high value of the novelty metric. Out of all the scenarios mentioned above, Case II gives the best result. In the next section, we will describe our proposed algorithm to overcome some of the pitfalls associated with the greedy cover algorithm and obtain better performance.

4. THE PROPOSED SIMILARITY-NOVELTY METRIC BASED ALGORITHM

Refinements to the greedy cover algorithm involve the computation of the novelty metric for the extended segment concerning the original coverage. The advertiser must establish a threshold for an acceptable level of novelty within an extended audience coverage. From the selected covers that meet the novelty threshold, the advertiser should proceed with the extended segment that boasts a considerable audience size and conversion rate, all achieved with an optimal cost per thousand impressions. Algorithm 2 comprehensively captures the novelty-similarity metric approach. Similar to the greedy cover algorithm, the initial cover consists of a segment of the population satisfying a broad spectrum of conditions based on a set of features. When the advertiser perceives that campaign requisites are unmet by the initial cover, they relax the stringent criteria applied to the features. This approach enables the similarity measure between the original and extended segments to attain its peak while simultaneously amassing a novel set of audience members to the population. It is important to note that there is a clear distinction that we observe in this approach. Unlike the greedy cover algorithm that returns a set of possible inventories, this approach returns a single extended audience segment that provides the best quantitative result to the advertiser. To test this method we will perform a case study of a sample Kaggle dataset which consists of conversion data from previous advertisement campaigns.

Algorithm 2 Novelty-Similarity metric based algorithm

1: **Input:** Initial cover $\{S\}$, Inventory Ω, User-defined Novelty threshold $0 \leq \alpha \leq 1$, required additional impressions N
2: **Output:** Optimal extended segment
3: **procedure** SIMILARITY-NOVELTY ALGORITHM(S, α)
4: Initialize variables: MaxImps=Imps(S), MaxCR=CR(S), MinCPTI=CPTI(S)
5: **for all** $S' \in \Omega$ **do**
6: **if** $\text{Nov}(S, S') > \alpha$ & $\text{Imps}(S') > N$ & $\text{CR}(S') > \text{MaxCR}$ & $\text{CPTI}(S') > \text{MaxCPTI}$ **then**
7: Update MaxImps
8: Update MaxCR
9: Update MinCPTI
10: **end if**

11: **end for**
12: **Return** S' with MaxSize, MaxCR, and MinCPTI
13: **end procedure**

5. CASE STUDY: KAGGLE CONVERSION DATA

To comprehend the effectiveness of the modifications applied to the conventional greedy approach algorithm, we will examine a practical implementation of the novelty-similarity-based approach described in the earlier section using the Kaggle conversion dataset. This dataset comprises advertisement campaigns with three target features: Age, Gender, and Interest. The advertiser chooses the initial coverage while imposing conditions on each constraint. We will generate six extended audience segments for every initial coverage by gradually relaxing one or multiple constraints. We shall then compare the initial coverage with the extended audience segments using a variety of performance metrics, ultimately determining the most favorable outcome among the extended segments. The metrics used to compare the performance of various extended segments are the number of impressions (Imps), cost incurred (Cost), cost per thousand impressions (CPTI), number of clicks the campaign garners (Clicks), total conversions (Conv), approved conversions (App), conversion rate (CR). We present three case studies that illustrate unique situations that surface with the changing novelty threshold and the difference in the overall performance of the various extended covers. In these cases, we primarily observe how the results obtained from the greedy cover algo- rithm differ from what the novelty-similarity metric-based algorithm generates. The three cases can be broadly described as:

- The first case study we present is the most commonly observed scenario where the choice of the best audience cover differs with the changing novelty threshold and these results are different from what the greedy cover algorithm provides.
- The second case study is the best possible outcome that any advertiser would want. Here, we will observe that the results from both approaches overlap, providing a segment with the largest audience size, maximum impressions, a reasonable cost, and a high conversion rate.
- The third case study is a specific subset of the first case study as it provides the same result for different novelty thresholds in the novelty-similarity approach but this result differs from the one provided by the greedy cover algorithm.

5.1. Case Study 1

We present in-depth information about our initial coverage, and we have identified six extended segments corresponding to this initial coverage, shown in Table 1. We have chosen the six extended segments in a way that maximizes the similarity metric between the original and the extended audience while at the same time incorporating new audience members into its fold. Table 2 shows a comparative analysis of the quantitative performance of various extended audience covers on a multiplicity of mathematical metrics. These are measured against the initial cover to surmise the best performance scenario at a reasonable cost and high conversion rate. Table 3 shows the value of the novelty metric for each of the extended audience covers with respect to the initial cover. To observe how the extended covers perform against the initial cover especially when it comes to the cost incurred for thousand impressions and conversion rate, we have used bar charts in data visualization in Figure 7.

Table 1: Audience segments for case study 1

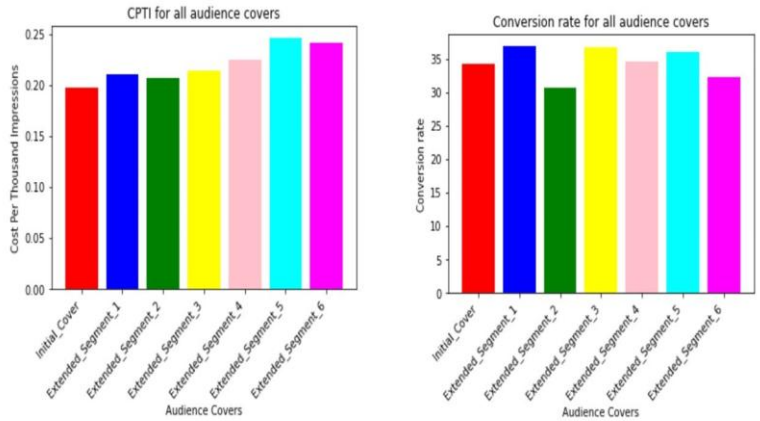
Audience Segment	Target Constraints
Initial Cover (Segment 0)	(Age=30-34 and Gender=M and Interest=15)
Extended Segment 1	(Age=30-34 and Gender=M)
Extended Segment 2	(Age=30-34 and Interest=15)
Extended Segment 3	(Gender=M and Interest=15)
Extended Segment 4	(Age=30-34)
Extended Segment 5	(Gender=M)
Extended Segment 6	(Interest=15)

Table 2: Performance Metric Analysis for case study 1

Segment	Imps	Cost	CPTI	Clicks	Conv	App	CR
0	4,846,178	956.90	0.197	510	108	37	34.26
1	36,421,443	7,640.92	0.209	4,384	812	299	36.82
2	6,515,339	1,348.51	0.206	772	147	45	30.61
3	7,002,876	1,496.33	0.213	847	131	48	36.64
4	67,993,019	15,252.40	0.224	9,483	1,431	494	34.52
5	98,571,981	24,202.61	0.245	14,287	1,620	584	36.04
6	10,745,856	2,597.26	0.241	1,609	195	63	32.30

Table 3: Calculating final quantitative measures for case study 1

Audience Segment	Novelty Score	Audience Size
1	0.93	229
2	0.34	23
3	0.50	30
4	0.96	426
5	0.97	592
6	0.70	51



(a) Cost per thousand impressions (b) Conversion Rate

Figure 7: Cost and Conversion Rates for case Study I

Table 4 captures the results obtained from the novelty-similarity metric and the greedy cover algorithm ($\alpha = 0$). The greedy cover algorithm gives segment 5 the best-extended audience segment as it encapsulates the maximum audience size. The novelty-similarity metric algorithm results in

Table 4: Comparison of Similarity-Novelty and Traditional Greedy Approach for Case Study 1

α	**Accepted Segments**	**Best Segment**	**Imps**	**CPTI**	**Conv Rate**
0 (Greedy)	1,2,3,4,5,6	5	98,571,981	0.245	36.04
0.5	1,3,4,5,6	1	36,421,443	0.209	36.82
0.7	1,4,5,6	1	36,421,443	0.209	36.82
0.9	1,4,5	1	36,421,443	0.209	36.82
0.95	4,5	4	67,993,019	0.224	34.50

changes with the values of α. At $\alpha = 0.95$ segment #1 did not qualify as it did not qualify the novelty threshold.

5.2. Case Study 2

We present in-depth information about our initial coverage, and we have identified six extended segments corresponding to this initial coverage shown in Table 5. We have chosen the six extended segments in a way that maximizes the similarity metric between the original and the extended audience while at the same time incorporating new audience members into its fold. Table 6 shows a comparative analysis of the quantitative performance of various extended audience covers on a multiplicity of mathematical metrics. These are measured against the initial cover to surmise the best performance scenario at a reasonable cost and high conversion rate. Table 7 shows the value of the novelty metric for each of the extended audience covers with respect to the initial cover. To observe how the extended covers perform against the initial cover, especially when it comes to cost incurred for thousand impressions and conversion rate, we have used bar charts in data visualization in Figure 8.

Table 5: Extended Audience Covers for case study 2

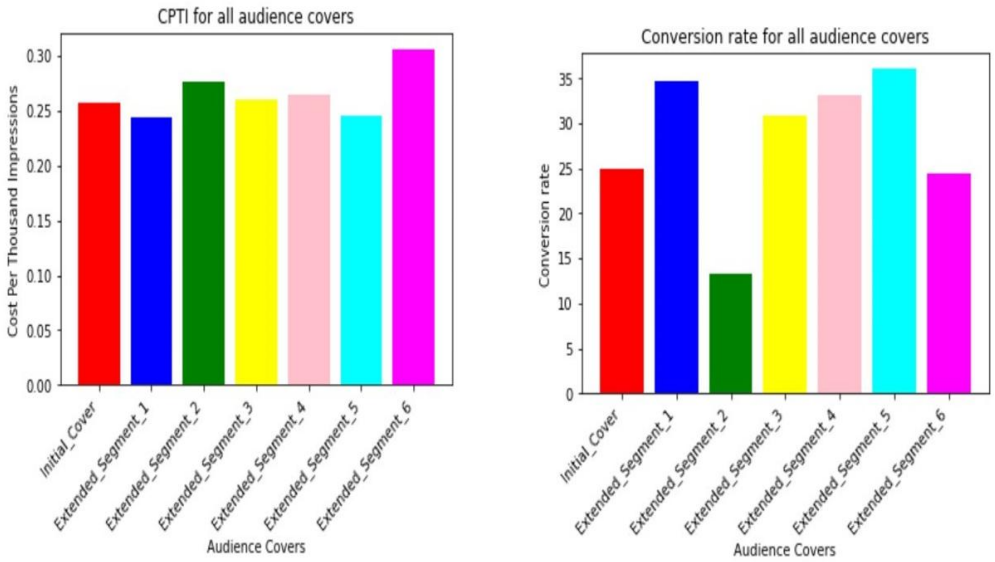
Extended Segment	**Target Constraints**
Initial Cover (Segment 0)	(Age=35-39, Gender=M, Interest=25)
Extended Segment 1	(Age=35-39 and Gender=M)
Extended Segment 2	(Age=35-39 and Interest=25)
Extended Segment 3	(Gender=M and Interest=25)
Extended Segment 4	(Age=35-39)
Extended Segment 5	(Gender=M)
Extended Segment 6	(Interest=25)

Table 6: Performance Metric Analysis for case study 2

Segment	Imps	Cost	CPTI	Clicks	Conv	App	CR
0	472,984	121.37	0.256	72	4	1	25.00
1	20,665,139	5,051.08	0.244	2,933	322	112	34.78
2	830,612	229.67	0.276	149	15	02	13.33
3	2,647,123	687.23	0.26	413	42	13	30.96
4	42,104,644	11,112.43	0.264	7,094	626	207	33.06
5	98,571,981	24,202.61	0.245	14,287	1,620	584	36.04
6	5,251,719	1,603.86	0.305	1,066	78	19	24.36

Table 7: Calculating final quantitative measures for case study 2

Audience Segment	Novelty Score	Audience Size
1	0.978	139
2	0.625	8
3	0.700	10
4	0.987	248
5	0.994	592
6	0.884	26



(a) Cost per thousand impressions

(b) Conversion Rate

Figure 8: Cost and Conversion Rates for case Study II

Finally, Table 8 shows the most desirable result for advertisers. In this case, the results obtained from the greedy cover algorithm and the novelty-similarity metric algorithm for every value of α

Table 8: Comparison of Similarity-Novelty and Traditional Greedy Approach for Case Study 2

α	Accepted Segments	Best Segment	Imps	CPTI	Conv Rate
0 (Greedy)	1,2,3,4,5,6	5	98,571,981	0.245	36.04
0.7	1,3,4,5,6	5	98,571,981	0.245	36.04
0.75	1,4,5,6	5	98,571,981	0.245	36.04
0.95	1,4,5	5	98,571,981	0.245	36.04
0.98	4,5	5	98,571,981	0.245	36.04

are the same. Here, audience segment #5 has the maximum audience size, highest conversion rate, highest impressions, and the highest novelty score at a reasonable cost per thousand impressions.

5.3. Case Study 3

We present in-depth information about our initial coverage and we have identified six extended segments corresponding to this initial coverage shown in Table 9. For each of these extended segments, we have computed the novelty metric and the final audience size in Table 10. Performance metrics, including cost per thousand impressions and conversion rate, are shown in Figure 9 and have been computed for both the initial audience coverage and the the extended segment is shown in Table 11.

Table 9: Extended Audience Covers for Case Study 3

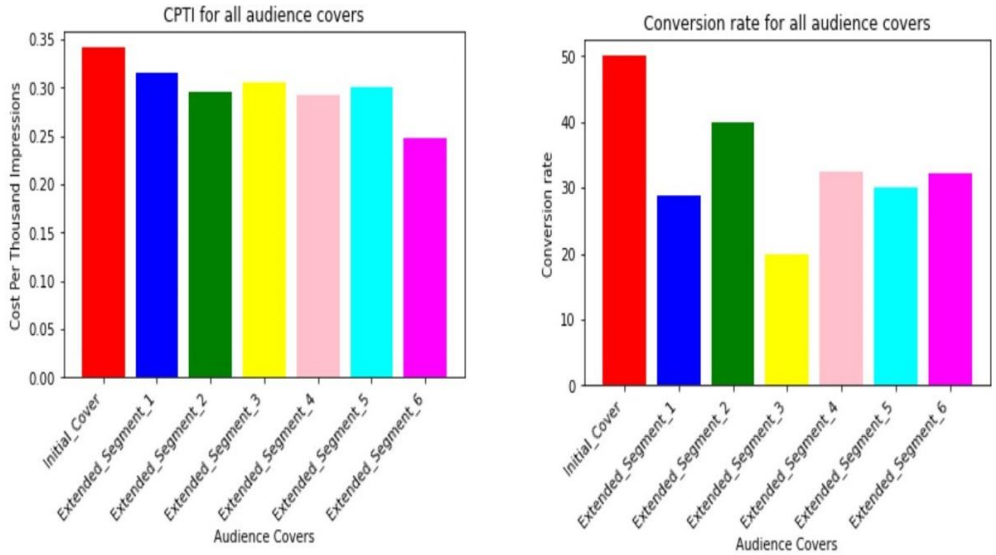
Extended Segment	Target Constraints
Initial Cover (Segment 0)	(Age=40-44, Gender=F, Interest=7)
Extended Segment 1	(Age=40-44 and Gender=F)
Extended Segment 2	(Age=40-44 and Interest=7)
Extended Segment 3	(Gender=F and Interest=7)
Extended Segment 4	(Age=40-44)
Extended Segment 5	(Gender=F)
Extended Segment 6	(Interest=7)

Table 10: Performance Metric Analysis for Case Study 3

Segment	Imps	Cost	CPTI	Clicks	Conv	App	CR
0	89,378	30.51	0.341	24	2	1	50
1	23,396,175	7,396.57	0.316	5,177	322	93	28.88
2	333,393	98.57	0.295	72	5	2	40
3	535,040	163.92	0.30	115	15	3	20
4	39,604,307	11,589.73	0.292	7,736	523	170	32.50
5	114,862,847	34,502.62	0.30	23,878	1,644	495	30.11
6	2,612,839	648.93	0.248	410	59	19	32.20

Table 11: Calculating final quantitative measures for Case Study 3

Audience Segment	Novelty Score	Audience Size
1	0.971	107
2	0.400	5
3	0.700	10
4	0.985	210
5	0.994	551
6	0.875	24



(a) Cost per thousand impressions

(b) Conversion Rate

Figure 9: Cost and Conversion Rates for case Study III

Table 12: Comparison of Similarity-Novelty and Traditional Greedy Approach for Case Study 3

α	Accepted Segments	Best Segment	Imps	CPTI	Conv Rate
0 (Greedy)	1,2,3,4,5,6	5	114,862,847	0.300	30.11
0.50	1,3,4,5,6	4	39,604,307	0.292	32.50
0.75	1,4,5,6	4	39,604,307	0.292	32.50
0.90	1,4,5	4	39,604,307	0.292	32.50
0.98	4,5	4	39,604,307	0.292	32.50

Finally, Table 12 shows the results of this case. In this example, the novelty-similarity metric-based algorithm gives the same result for all values of α due to its high novelty value. However, the results differ from those obtained through the greedy cover algorithm.

In summary, we have explored three broad situations that encapsulate many different results that we may get when we apply the greedy cover algorithm and the novelty-similarity metric-based algorithm with varying thresholds to a given advertisement campaign. The first instance is a general case where the results of both algorithms are different, and these change with the threshold for the novelty-similarity metric. The second instance is the most favorable outcome for the advertiser, where he gets the best aspects from both the algorithms. The third instance is a specific outcome described as a subset of the first. Although the results do not vary within the varying values of the novelty threshold, they differ from the one given by the greedy cover algorithm.

6. CONCLUSION

This investigation has introduced a principled enhancement to the classical greedy cover algorithm for audience extension by integrating rigorously defined similarity and novelty metrics. The resultant framework subsumes the conventional greedy approach as a limiting case but significantly extends its utility by enabling a more discriminating selection of extended audience segments. Through a series of empirically grounded case studies, we have demonstrated that the proposed algorithm affords advertisers a quantitatively and qualitatively superior mechanism for audience amplification, with demonstrable gains in conversion efficacy and cost-efficiency. Looking forward, we envision that the similarity-novelty paradigm may serve as a foundational construct for broader applications, particularly in collaborative filtering and recommender systems. Its integration could endow advertising platforms with an expanded repertoire of machine learning tools capable of producing bespoke audience recommendations, thus advancing the frontier of data-driven marketing strategy.

DECLARATIONS

CONFLICT OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this work.

FUNDING

This research was conducted without any external funding.

ACKNOWLEDGEMENTS

We thank the Department of Computer Science at Boston University for their support.

DATA AVAILABILITY

All the relevant data and analysis are available via:
https://drive.google.com/?????

REFERENCES

[1] J. Smith. Changing trends in advertising effectiveness. *Journal of Marketing Trends*, 45(2):34– 47, 2020.

[2] A. Jones. The effectiveness of traditional advertising approaches. *Marketing Quarterly*, 62(4):18–29, 2018.

[3] V. Gallego, J. Lingan, A. Freixes, A. Juan, and C. Osorio. Applying machine learning in marketing: An analysis using the nmf and k-means algorithms. *Information*, 15(7), 2024.

[4] D. Dwivedi and et al. Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, 59:102168, 2021.

[5] Choi. J. and K. Lim. Identifying machine learning techniques for classification of target ad-vertising. *ICT Express*, 6(3):175–180, 2020.

[6] D. Herhausen, S.F. Bernritter, E. Ngai, A. Kumar, and D. Delen. Machine learning in market- ing: Recent progress and future research directions. *Journal of Business Research*, 170:114254, 2024.

[7] M.S. Ullal, I.T. Hawaldar, R. Soni, and M. Nadeem. The role of machine learning in digital marketing. *SAGE Open*, 11(4):21582440211050394, 2021.

[8] R. Wilson. Digital advertising trends: Precise audience targeting. *Advertising Insights*, 9(1):56– 68, 2022.

[9] D. Martinez. The role of data management platforms in modern advertising. *Data Marketing Journal*, 35(3):82–95, 2021.

[10] B. Chen and C. Smith. Data-driven insights for audience targeting strategies. *Journal of Advertising Analytics*, 28(2):123–138, 2020.

[11] I. Ahmadi and et. al. Overwhelming targeting options: Selecting audience segments for online advertising. *International Journal of Research in Marketing*, 41(1):24–40, 2024.

[12] C. Gromit Yeuk-Yin and et al. Interactive audience expansion on large scale online visitor data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2621–2631, New York, NY, USA, 2021. Association for Computing Machinery.

[13] H. Liu, D. Pardoe, K. Liu, M. Thakur, F. Cao, and C. Li. Audience expansion for online social network advertising. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[14] E. Harris. Audience segmentation challenges in modern advertising. *Marketing Challenges*, 51(4):201–215, 2023.

[15] M. Anderson. Metrics for evaluating audience reach in digital advertising. *Ad Metrics Review*, 14(3):76–89, 2022.

[16] P. Williams and R. Davis. Audience extension strategies: Challenges and opportunities in digital advertising. *Journal of Marketing Insights*, 15(1):45–58, 2023.

[17] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2016.

[18] T. Hastle. *Elements of Statistical Learning*. Pearson, 2018.

[19] L. Chen and K. Anderson. Optimizing the greedy approach algorithm for audience extension. *Journal of Advertising Research*, 22(3):176–192, 2023.

[20] L. Brown. Algorithms for audience extension in digital advertising. *Algorithmic Insights*, 7(4):189–202, 2021.

[21] C. L. Johnson and D. W. White. Enhancing audience segmentation in digital advertising: A comparative analysis of algorithms. *International Journal of Advertising*, 30(4):205–220, 2022.

[22] R. Miller. Algorithmic challenges in audience extension. *Algorithmic Trends*, 18(3):76–89, 2019.

[23] S. Rogers. Algorithmic pitfalls in audience extension. *Algorithm Analysis*, 25(1):45–57, 2020.

[24] Q. Wang. Enhancements for greedy approach algorithm in audience extension. *Algorithmic Innovations*, 11(2):87–101, 2023.

[25] J. Lee. Algorithmic modifications for improved audience targeting. *Journal of Advertising Optimization*, 40(3):142–155, 2021.

[26] J. Shen and et. al. Effective audience extension in online advertising. *Journal of the Association for Computing Machinery*, page 2099–2108, 2015.

[27] A. Brown and B. Jones. Data management platforms: A comprehensive overview of features and applications. *Journal of Advertising Technology*, 18(2):67–84, 2021.

[28] J. R. Smith and M. S. Johnson. The evolution of advertising techniques: From traditional broadcasting to digital targeting. *Journal of Marketing Research*, 56(3):387–402, 2019.

[29] S. Pattnaik and E. Pinsky. α-based similarity metric in computational advertising: A new ap- proach to audience extension. In *Proceedings of the EAI International Conference on Computer Science, Engineering & Communication Systems (CSECS 2023)*, June 2023.