人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第35回)
SIG-AM-35-01

LLMを用いた推薦リスト生成のための ペアワイズ比較・集約手法についての予備的検討

Preliminary Study on Pairwise Comparison and Aggregation Methods for Recommendation List Generation Using LLMs

谷 知拓[1*]　柴田 祐樹[2]　高間 康史[2]
Tomohiro Tani[1]　Hiroki Shibata[2]　Yasufumi Takama[2]

[1] 東京都立大学 システムデザイン学部
[1] Faculty of Systems Design, Tokyo Metropolitan University
[2] 東京都立大学大学院 システムデザイン研究科
[2] Graduate School of Systems Design, Tokyo Metropolitan University

Abstract: This paper investigates pairwise comparison and aggregation approaches for generating recommendation lists using Large Language Models (LLMs) in a zero-shot and scalable manner. When dealing with a large number of candidate items that cannot be evaluated by LLMs in a single prompt, it becomes necessary to divide the task into multiple prompts and aggregate the results. The proposed method prompts LLMs to estimate preference order between two items, and generates recommendation lists through Bradley-Terry-Luce (BTL) aggregation. This paper reports the results of preliminary experiments and discusses the impact of the number of item pairs and LLM input batch size on the consistency and accuracy of recommendations.

1 はじめに

本稿では，LLM を用いたゼロショット・スケーラブルな推薦リスト生成手段として，ペアワイズ比較を集約するアプローチに着目し，予備実験を行った結果について報告する．

近年，大規模言語モデル（Large Language Model: LLM）の急速な発展により，自然言語処理の様々なタスクにおいて革新的な成果が得られている．推薦システムの分野においても，LLM を活用した新たなアプローチが注目を集めており，従来の協調フィルタリングや内容ベースフィルタリングとは異なる観点から，ユーザの嗜好を理解し推薦を生成する手法が提案されている [2, 3, 4, 5]．特に，LLM の持つゼロショット学習能力を活用することで，事前の学習データや特徴量エンジニアリングを必要とせずに，テキスト記述のみから推薦を生成できる可能性が示されている [1]．

しかしながら，LLM を用いた推薦システムの実用化においては，いくつかの技術的課題が存在する．特に，推薦候補となるアイテム数が多い場合，すべてのアイテムを一度に LLM に入力することは，入力トークン数の制限により困難である．この問題に対処するため，候補アイテムを複数のバッチに分割し，段階的に処理を行うアプローチが考えられるが，その際に候補アイテムのバッチへの分割，LLM の出力の統合をどのように行い，最終的な推薦リストを生成するかが重要な課題となる．

本稿では，この課題に対するアプローチとして，ペアワイズ比較に基づく手法に着目する．具体的には，アイテムのペアごとに LLM に選好関係を評価させ，得られた比較結果を Bradley-Terry-Luce（BTL）モデル [8] を用いて集約することで，全体的な推薦リストを生成する手法を提案する．ペアワイズ比較アプローチは，各比較において考慮すべき情報量を限定できるため，LLM の入力制限に対して頑健であり，また比較の並列処理が可能であるという利点を持つ．

本稿では，提案手法の予備的な検討として，ペア数および LLM への入力バッチサイズが推薦の一貫性と精度に与える影響について評価実験を行った結果を報告する．実験では，実際の商品レビューデータを用いて，異なる設定での推薦リスト生成を行い，その性能を比較分析する．実験の結果，ペア数が増加するにつれて推薦精度と一貫性が向上する傾向が確認されたが，50 ペアを超えると性能が飽和することが観測された．

*連絡先：東京都立大学システムデザイン学部
〒 191-0065 日野市旭が丘 6-6
E-mail: tani-tomohiro@ed.tmu.ac.jp

また，バッチサイズについては，バッチサイズ 2 の純粋なペアワイズ比較において最も高い一貫性が得られたことを報告する．

2 関連研究

2.1 LLM を用いた推薦システム

大規模言語モデルを推薦システムに応用する研究は，近年活発に進められている．特に，事前の学習データなしに推薦を行うことが可能なゼロショット学習による推薦生成 [1] は，コールドスタート問題の解決やビックデータを必要としない柔軟な推薦システムの構築手段として，期待が寄せられている．また，協調フィルタリングの概念を LLM に組み込んだ手法 [2, 6] や，グラフ構造を活用した手法 [3]，生成的アプローチによる個人化推薦 [4, 5] など，様々な観点からの研究が進められている．

しかし，既存の LLM ベース推薦手法の多くは，候補アイテム数が限定的な場合を想定しており，大規模なアイテム集合に対するスケーラビリティの課題が残されている [7].

2.2 ペアワイズ比較に基づく選好学習

2 つの選択肢を比較するペアワイズ比較は，多数の選択肢を同時に評価するよりも認知負荷が低く，より正確な判断が可能であることが知られている [9]．LLM においても，ペアワイズ比較は有効なアプローチとして注目されている．複数のアイテムを一度に評価させる場合と比較して，2 つのアイテムの相対的な優劣を判断させることで，より一貫性のある選好情報を獲得できることが報告されている [1, 7].

Bradley-Terry-Luce（BTL）モデルは，ペアワイズ比較結果から全体的なランキングを推定する統計モデルである [8]．各アイテムに潜在的な強度パラメータを仮定し，アイテム i がアイテム j より選好される確率を以下のように表現する：

$$P(i \succ j) = \frac{\pi_i}{\pi_i + \pi_j} \tag{1}$$

ここで，π_i はアイテム i の強度パラメータであり，観測された比較結果から最尤推定や反復アルゴリズムにより求められる．例えば，3 つのアイテム A，B，C に対して「A が B に勝つ」「B が C に勝つ」「A が C に勝つ」という比較結果が得られた場合，これらの観測確率を最大化するように各アイテムの強度パラメータ π_A, π_B, π_C を推定し，その値に基づいて全体のランキング（この例では A > B > C）を決定する．

BTL モデルは，トーナメントやスポーツランキング等の分野で広く利用されている．

3 提案手法

3.1 システム概要

本稿では，LLM を用いたスケーラブルな推薦システムを実現するため，ペアワイズ比較と BTL モデルによる集約に基づく推薦システムを提案する．提案システムは，以下の 3 つの主要コンポーネントから構成される：(1) 候補アイテムのペアワイズ比較を行う LLM モジュール，(2) 比較結果を集約する BTL モデル，(3) 最終的な推薦リストを生成するランキングモジュール．

システムの処理フローは以下の通りである．まず，ユーザの嗜好情報と候補アイテム集合を入力として受け取る．次に，候補アイテム集合からアイテムペアを生成し，各ペアについて LLM に選好判定を要求する．LLM は，ユーザの嗜好情報に基づいて，どちらのアイテムがより適切かを判定する．得られた比較結果は BTL モデルに入力され，各アイテムの強度パラメータが推定される．最後に，推定された強度パラメータに基づいてアイテムをソートし，推薦リストを生成する．

本システムの特徴は，アイテム数が多い場合でも，ペアワイズ比較により処理を分割できる点にある．すべてのアイテムを一度に LLM に入力する方法と比較して，本システムでは各比較を独立に実行できるため，並列処理が可能であり，スケーラビリティが向上する．

3.2 ペアワイズ比較の実施

ペアワイズ比較では，候補アイテム集合から 2 つのアイテムを選択し，LLM に対してどちらがユーザにとって適切かを判定させる．具体的には，以下の形式で LLM にプロンプトを与える：

> ユーザの嗜好情報：[ユーザの過去の購買履歴やレビュー]
>
> 以下の 2 つのアイテムのうち，このユーザにより適していると思われるものを選択してください．
>
> アイテム A：[アイテム A の特徴・説明]
>
> アイテム B：[アイテム B の特徴・説明]

ペアの生成方法については，候補アイテム集合から n 個のペアを，重複して同じペアが選ばれないようにサンプリングする．ペア数 n は重要なパラメータであり，多すぎると LLM へのクエリ数が増加しコストが

上昇する一方，少なすぎると推薦精度が低下する可能性がある．

バッチサイズは，一度の LLM クエリで比較するアイテム数を制御する．$b = 2$ の場合は純粋なペアワイズ比較となり，$b > 2$ の場合は複数ペアを統合したアイテム群での比較となる．異なるバッチ間では LLM の判断に矛盾が生じる可能性がある．BTL モデルはペア間で選好順序に矛盾がある場合も処理可能であるが，バッチサイズは推薦の一貫性に影響する重要なパラメータと考える．

3.3 BTL 集約による推薦リスト生成

LLM によるペアワイズ比較結果を集約するため，BTL モデルを使用する．本稿では計算効率と数値安定性の観点から，対数スコア $s_i = \log \pi_i$ を用いた定式化を採用する．この関係を式 (1) に代入すると，次式が得られる：

$$P(i \succ j) = \frac{1}{1 + \exp(s_j - s_i)} \tag{2}$$

ロジスティック関数を用いた表現により，スコア差が大きいほど選好確率が 1 に近づき，また勾配に基づく最適化が容易になる．

3.3.1 パラメータ推定アルゴリズム

本研究の実装では，確率的勾配降下法（SGD）を用いてスコアパラメータを推定する．観測されたペアワイズ比較結果の集合を $\mathcal{D} = \{(w_k, l_k)\}$ とする．ここで，w_k は k 番目の比較における勝者，l_k は敗者を表す．

各比較 (w_k, l_k) に対して，以下の更新式でスコアを調整する：

$$p_k = P(w_k \succ l_k) = \frac{1}{1 + \exp(s_{l_k} - s_{w_k})} \tag{3}$$

$$s_{w_k} \leftarrow s_{w_k} + \alpha(1 - p_k) \tag{4}$$

$$s_{l_k} \leftarrow s_{l_k} - \alpha(1 - p_k) \tag{5}$$

ここで，α は学習率（デフォルト値：0.01）である．この更新により，勝者のスコアは増加し，敗者のスコアは減少する．更新量 $(1 - p_k)$ は，現在のモデルによる予測確率と実際の結果（勝者の勝利確率=1）との差に相当する．

3.3.2 不確実性の定量化

BTL モデルでは，アイテムペア (i, j) の比較における不確実性を以下のように定量化できる：

$$U(i, j) = 2 \times \min\{P(i \succ j), P(j \succ i)\} \tag{6}$$

この不確実性指標は，両アイテムのスコアが近い場合に最大値 1 をとり，スコア差が大きい場合に 0 に近づく．本システムでは，この不確実性を利用してアクティブサンプリングを行い，不確実性の高いペアを優先的に LLM に評価させることで，クエリ効率を向上させている．

3.3.3 収束性と計算量

提案手法では，全比較データに対して 100 回の反復を行うことで収束を図る．1 回あたりの計算量は $O(|\mathcal{D}|)$ であり，全体の計算量は $O(100 \times |\mathcal{D}|)$ となる．実験では，$|\mathcal{D}|$ は最大 100 ペアであり，計算は数ミリ秒で完了した．

最終的に，推定されたスコア $\{s_i\}$ に基づいてアイテムを降順にソートすることで，推薦リストを生成する．この手法により，限られた数のペアワイズ比較から，全アイテムの相対的な順位を効率的に推定することが可能となる．

4 評価実験

4.1 データセット

Amazon Review Dataset 2023[1][7] の Movies and TV カテゴリを使用して評価実験を行った．このデータセットは，Amazon プラットフォーム上で公開されている映画・TV 番組に関するユーザレビューを含んでおり，推薦システムの評価に広く利用されている．Movies and TV カテゴリを選択した理由は，LLM が事前学習において映画や TV 番組に関する豊富な知識を獲得していることが予想され，アイテムの内容を理解した上での推薦が期待できるためである．

データセットから，各ユーザに対して正解（実際に高評価を付けたアイテム）1 件と，ランダムに選択した 19 件の計 20 件を候補アイテムとして使用した．この設定により，推薦タスクは 20 個のアイテムから最適なものを識別する問題となる．候補アイテム数を 20 件に限定した理由は，ペアワイズ比較の効果を明確に評価するためと，実験の計算コストを現実的な範囲に抑えるためである．

各アイテムについては，タイトル，ジャンル（カテゴリ），平均評価およびレビュー件数，特徴，製品説明等のメタデータを使用し，LLM がアイテムの内容を理解できるようにした．ユーザの嗜好情報としては，最

[1] https://amazon-reviews-2023.github.io/

近評価したアイテム（2-5 件程度）のメタデータを要約したテキストを使用し，各ユーザの好みを LLM に伝えられるようにした．

4.2 実験設定

4.2.1 評価指標

推薦システムの性能を多角的に評価するため，以下の指標を使用する．

(1) 推薦精度指標

- **Hit@10**：正解アイテムが推薦リストの上位 10 位以内に含まれる割合．この指標は推薦システムが関連アイテムを上位に配置できているかを評価する．
- **平均逆順位（Mean Reciprocal Rank, MRR）**：正解アイテムの順位の逆数の平均値．より上位に正解が現れるほど高い値となる．
- **平均順位**：正解アイテムが推薦リスト中に現れる順位の平均値．20 個の候補中での絶対的な位置を示す．

(2) 推薦の一貫性指標

同一ユーザ・同一条件で複数回（5 回）推薦を行った際の結果の安定性を以下の指標で評価する：

- **Jaccard 類似度**：2 つの推薦リストの上位 10 アイテム集合に対する類似度．$J(A,B) = |A \cap B|/|A \cup B|$ で計算され，0 から 1 の値をとる．
- **Spearman 順位相関係数**：2 つの推薦リストの上位 10 アイテムのうち，共通するアイテムの順位相関．順位の一致度を-1 から 1 の範囲で評価する．

4.2.2 バッチサイズに関する実験

バッチサイズ b を $\{2, 5, 10\}$ と変化させて得られたペアの集合から，ランダムに 50 ペアを抽出し，推薦の一貫性に与える影響を評価した．バッチサイズは LLM への一度の入力で比較するアイテム数を制御するパラメータであり，b 個のアイテムから $\binom{b}{2}$ 個のペアワイズ比較が生成される．

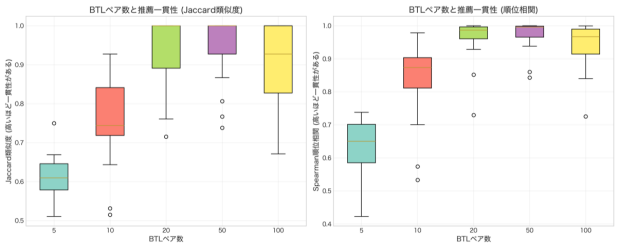

図 1: ペア数と推薦の一貫性の関係．(a) Jaccard 類似度，(b) Spearman 順位相関係数．

4.2.3 実験手順

各実験は以下の手順で実施した：

1. 推薦対象（U）の各ユーザについて正解 1 件＋ランダム 19 件の候補アイテムを準備．
2. アクティブサンプリングにより，不確実性（式 (6)）の高いアイテム集合（サイズ=b）を選択．
3. 選択されたアイテム集合からアイテムペアを生成して LLM に選好順序を判定させ，3.3 節で述べた手順により，BTL モデル（反復 100 回）により各アイテムの s_i を推定．
4. s_i に基づいてアイテムをソートし，推薦リストを生成．
5. 評価指標を計算．

すべての実験条件で同一のユーザ・アイテム集合を使用することで，公平な比較を実現した．

4.3 実験結果

4.3.1 ペア数が精度・一貫性に与える影響

BTL モデルに入力するペア数が推薦性能に与える影響を評価するため，ペア数を $\{5, 10, 20, 50, 100\}$ の 5 段階で変化させて実験を行った．バッチサイズは 2 に固定し，50 ユーザに対して各条件で 5 回の試行を実施した．

表 1 に，ペア数と推薦精度の関係を示す．Hit@10 は，ペア数の増加に伴って向上する傾向が観察されたが，ペア数 100 ではわずかに低下する傾向が見られた．同様に，平均逆順位（MRR）と平均順位においても，ペア数 50 付近で最良の性能を示し，それ以上では性能が飽和または低下する傾向が確認された．

図 1 に，同一条件での 5 回の試行結果に対する Jaccard 類似度と Spearman 順位相関係数を示す．両指標ともに，ペア数の増加に伴って一貫性が向上する傾向が観察された．特に Spearman 順位相関係数では，より滑

表 1: ペア数と推薦精度指標の統計サマリ（50 ユーザの平均値±標準偏差）

ペア数	Hit@10	MRR	平均順位
5	0.42 ± 0.18	0.18 ± 0.11	8.2 ± 3.4
10	0.58 ± 0.16	0.27 ± 0.13	6.5 ± 2.8
20	0.65 ± 0.14	0.33 ± 0.12	5.8 ± 2.5
50	0.72 ± 0.12	0.38 ± 0.11	5.2 ± 2.2
100	0.70 ± 0.13	0.36 ± 0.12	5.4 ± 2.3

表 2: ペア数と推薦の一貫性指標の統計サマリ（50 ユーザの平均値±標準偏差）

ペア数	Jaccard	Spearman
5	0.35 ± 0.12	0.42 ± 0.15
10	0.48 ± 0.10	0.56 ± 0.12
20	0.58 ± 0.09	0.65 ± 0.10
50	0.68 ± 0.08	0.73 ± 0.09
100	0.66 ± 0.09	0.71 ± 0.10

らかな上昇曲線が得られた．これは，順位情報を考慮した指標であるため，推薦リストの順位の安定性をより適切に評価できているためと考える．

ペア数が 50 を超えると，一貫性の低下が観察された．この現象は，過度に多くのペアワイズ比較を行うことで，ノイズや矛盾が蓄積される「過学習」に類似した現象が生じている可能性を示唆しており，推薦精度の低下の一因になったと考える．

4.3.2 バッチサイズが精度・一貫性に与える影響

次に，BTL に入力するペア数を 50 に固定した状態で，バッチサイズを変化させた際の影響を分析した．バッチサイズは，一度のプロンプトで比較するアイテム数を制御するパラメータであり，計算効率と推薦品質のトレードオフを決定する重要な要素である．

図 2 に，バッチサイズと推薦の一貫性指標の関係を示す．バッチサイズが 2 の場合に最も高い一貫性を示し，バッチサイズの増加に伴って一貫性が低下する傾向が観察された．

この結果は，小さなバッチサイズでは LLM がより明確な選好判断を行えることを示唆している．2 つのアイテムの直接比較は最もシンプルなタスクであり，LLM は一貫した判断を下しやすい．一方，多数のアイテムを同時に順位付けする必要がある場合，タスクの複雑性が増し，出力の変動が大きくなると考える．

また，正解アイテムの順位の標準偏差およびレンジ（最大値と最小値の差）（図 2 下段）も，バッチサイズの増加とともに単調に増加する傾向を示し，推薦結果の不安定性が顕著となった．

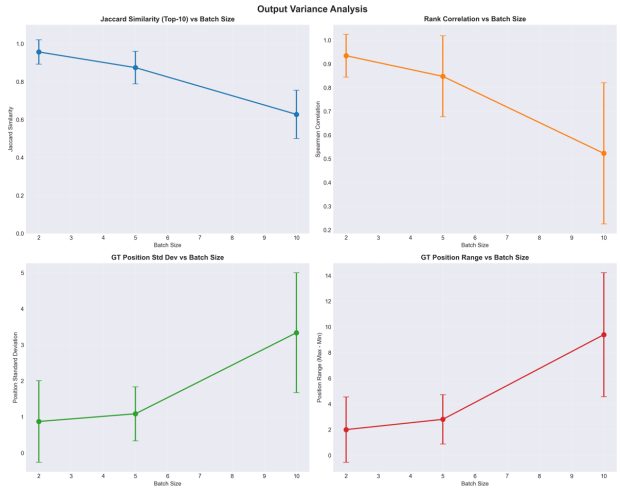

図 2: バッチサイズと推薦の一貫性の関係．上段左：Jaccard 類似度，上段右：Spearman 順位相関，下段左：正解順位の標準偏差，下段右：正解順位のレンジ（最大値-最小値）．エラーバーはユーザ間のばらつき（標準偏差）を示す．

4.4 考察

4.4.1 実験結果の解釈

本実験の結果から，LLM を用いたペアワイズ比較に基づく推薦システムにおいて，ペア数とバッチサイズが推薦品質に与える影響について，以下の重要な知見が得られた．

第一に，ペア数と推薦精度の関係において，単純な比例関係ではなく，ある閾値（本実験では 50 ペア）で性能が飽和する傾向が確認された．これは，BTL モデルが一定数以上の比較データから十分な統計的情報を抽出できることを示している．一方で，ペア数が 100 を超えると精度が若干低下する現象は，推移律違反の増加と関連していると考える．推移律違反などの矛盾した判断については今後詳細に分析する予定であるが，LLM の出力に内在する確率的な揺らぎが，大規模な比較集合において矛盾として顕在化し，BTL 集約の品質を低下させている可能性がある．

第二に，バッチサイズと推薦の一貫性の間には負の相関が観察された．バッチサイズ 2（純粋なペアワイズ比較）において最も高い一貫性が得られた理由として，タスクの認知的複雑性が考えられる．人間の意思決定研究においても，選択肢が増加すると判断の一貫性が低下することが知られており，LLM も同様の傾向を示すことが示唆される．また，大きなバッチサイズでは，アイテム間の相対的な特徴の差異が希薄化し，順位付けの基準が不安定になる可能性がある．

5 おわりに

本稿では，LLM を用いたゼロショット・スケーラブルな推薦リスト生成のために，ペアワイズ比較結果を BTL モデルで集約する手法を検討した．

評価実験では，精度指標（Hit@10，MRR，平均順位）と一貫性指標（Jaccard 類似度，Spearman 順位相関）がペア数，バッチサイズによりどのように変化するかを調査した．その結果，ペア数が増加するにつれて推薦精度と一貫性が向上する傾向が確認されたが，50 ペアを超えると性能が飽和することが観測された．また，バッチサイズ 2 において最も高い一貫性が得られた．

今後は，推移率違反などの矛盾の発生について調査する他，既存の協調フィルタリングやコンテンツベースフィルタリング手法との精度や計算効率の比較検証を進め，LLM を用いたゼロショット推薦手法の有効性を調査する予定である．

謝辞

本研究の一部は，JSPS 科研費 22K19836，23K24953 の助成を受けたものです．

参考文献

[1] Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., Zhao, W.X.: Large Language Models for Zero-Shot Recommender Systems, *ECIR2024*, pp. 364-381 (2023)

[2] Yao, S., Wu, L., Guo, Q., Hong, L., Li, J.: Collaborative Large Language Model for Recommender Systems, *WWW'24*, pp. 3162–3172 (2024)

[3] Wei, W., Ren, X., Tang, J., Wang, Q., Su, L., Cheng, S., Wang, J., Yin, D., Huang, C.: LLMRec: Large Language Models with Graph Augmentation for Recommendation, *WSDM'24*, pp. 806–815 (2024)

[4] Li, J., Zhang, W., Wang, T., Xiong, G., Lu, A., Medioni. G.: GPT4Rec: A Generative Framework for Personalized Recommendation and User Interests Interpretation, *SIGIR eCom 2023* (2023)

[5] Ngo, H., Nguyen, D.Q.: RecGPT: Generative Pre-training for Text-based Recommendation, *ACL2024*, pp. 302–313 (2024)

[6] Kim, S., Kang, H., Choi, S., Kim, D., Yang, M., Park, C.: Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System, *KDD'24*, pp. 1395–1406 (2024)

[7] Hou, Y., Li, J., He, Z., Yan, A., Chen, X., McAuley, J.: Bridging Language and Items for Retrieval and Recommendation, *arXiv preprint*, arXiv:2403.03952 (2024)

[8] Bradley, R.A., Terry, M.E.: Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons, *Biometrika*, Vol. 39, No. 3/4, pp. 324–345 (1952)

[9] Guo, S., Sanner, S.: Real-time Multiattribute Bayesian Preference Elicitation with Pairwise Comparison Queries, *AISTATS'2010*, pp. 289-296 (2010)