レシート印字名に基づく インストア商品の自動分類技術

Automated In-Store Product Classification Using Receipt Printing Information to Precisely Analyze Purchase Data

商品の分類作業を省人化し，レシートデータの分析による商品開発やマーケティング活動の活性化に貢献

近年，購買データを活用した市場分析やマーケティング高度化に対する需要が高まっています。しかし，総菜や生鮮食品などの店内で加工されるインストア商品は，国内共通の商品バーコードであるJAN（Japanese Article Number）コードを持たず，店舗ごとに独自のインストアコードとレシート印字名で管理されるため，購買データの分類作業が難しいという問題があります。そこで，レシート印字名から商品分類名をAIで自動推定する技術を開発しました。LightGMAIC（Light Graph-based Multi-Angled Item Categorization）による文字列の部分構造化と，レシート表記に特化したLLM（大規模言語モデル）を組み合わせ，87％以上の分類精度を実現できました。

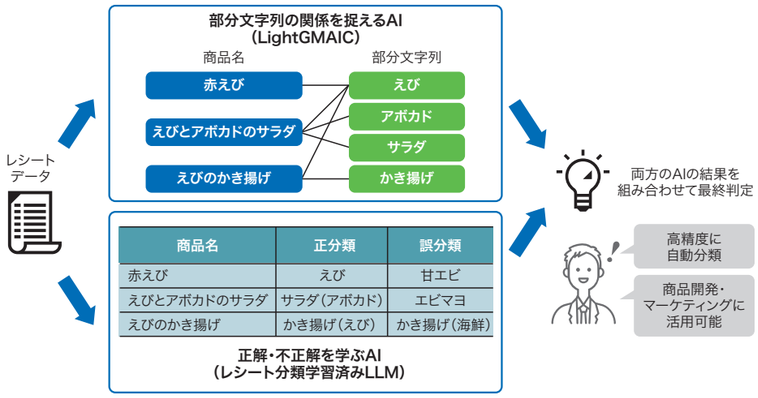

図1．2種類のAIを開発して組み合わせた自動分類技術の概要
部分文字列を用いたGNNとレシート分類学習を行ったLLMの2種類のAIを組み合わせたことで，高精度な分類を実現しました。

はじめに

レシートは，消費者の購買行動を詳細に記録する情報源としての重要性が増しています。東芝グループの電子レシートサービス“スマートレシート”のような電子レシートサービスが普及する中，複数の店舗の購買データを横断的に解析することで，消費者の行動を多角的に理解し，マーケティング施策や商品開発へ活用する需要が高まっています。しかし，これらの購買データを分析するには，商品を共通の分類体系に統合する必要があります。インストア商品は，その店舗でだけ有効な独自の商品バーコード（インストアコード）が設定されていて，国内共通のJANコードは設定されておらず，分類の統合が極めて困難でした。

背景と課題

購買データを適切に分析するには，商品が統合した分類体系にマッピングされる必要があります。分類が整備されていれば，ジャンル別売上分析や，類似商品の購買傾向把握，市場全体における商品の位置付け，自社商品と競合商品との相対比較などが，容易になります。また，小売事業者やメーカーがマーケティング施策を検討する際にも，分類情報は不可欠です。しかし，インストア商品はJANコードを持たないため外部データベースに登録されておらず，対応付けができないことから，小売事業者や分析事業者が扱いづらいことへの対策が課題でした。

更に，インストア商品の分類を難しくしている要因として，表記揺れの多さが挙げられます。同一分類の商品であっても，略語や独自省略も頻繁に用いられ，「海老アボカドサラダ」，「エビアボサラ」，「海老アボサラ」などの複数の派生表記が存在します。また，商品名に量目（「100 g，1パック」など）や加工方法（「レンジ可，タレ付」など）といった情報が付加されることも多く，分類に寄与しない情報と寄与する情報を区別する必要があります。このように，インストア商品は標準化されていないため，表記の揺れとノイズが分類精度を低下させる主要因となっていることへの対策が課題でした。

これらの課題を解決するには，単純な文字列一致に頼るだけでは不十分です。表記揺れにも強く，略語を含む文字列から意味的構造を導き出す仕組みが必要でした。また，新規に登場する商品にも柔軟に対応できる推論能力が求められました。そこで，文字列の構造を解析するGNN（Graph Neural Network）と，自然言語理解に優れたLLMを組み合わせることで，この課題に対処しました[1]。

自動分類技術の概要，及び実データを用いた評価

開発した自動分類技術は，文字列の部分構造を抽出してグラフ化するGNNモデル（LightGMAIC）と，レシート表記に特化してチューニングしたLLMを組み合わせたハイブリッド分類技術です（**図1**）。LLMの自然言語理解能力を活用し，商品名に込められた意味やニュアンスから分類を推論することが可能です。略語や，省略，店舗独自の命名規則などの多様な表記に対しても，文脈や語彙の意味関係を理解し，類似性や関連性を高精度で推定します。

従来のLLMは，一般的な言語知識に基づいて学習されているため，レシート特有の表記や業界固有の略語については十分な知識を持っていないという問題がありました。例えば，「エビアボサラ」や「タレ付」など，レシートに頻出する短縮表現や加工情報は，通常の言語モデルでは意味解釈が難しい場合があります。

これに対し，開発したLightGMAICは，商品名に含まれる特徴語を部分文字列単位で抽出し，それらの共起関係（ここでは，異なる商品名に共通した部分文字列が出現すること）や文字列位置情報をグラフ化して統合的に扱います。これにより，略語や省略表現が含まれる場合でも，文字列全体の意味構造を捉えられます。更に，GNNによるグラフ学習により，店舗ごとの表記揺れや略語の違いといった現場特有のパターンも学習できるため，従来のルールベースや辞書ベースでは対応が難しかったケースにも対応可能です。

最終的な分類結果は，LightGMAICとLLMの推論を統合することで導き出します。両モデルは，事前に教師データを用いて学習させており，文字列構造や表記揺れのパターンを適切に分類できるように内部パラメーターを調整しています。運用時には，分類結果を自動的に統合し，両方の推論が一致しないケースや信頼度が低いケースだけを人手で確認する仕組みを採用することで，作業効率を大幅に向上させることが可能です。この方式により，レシート特有の複雑な表記や，略語，店舗独自の記述などに対応しながら，高精度かつ安定した分類を実現できます。また，確認作業を重点化することで，運用コストを抑えながら，安定した分類サービスを継続的に提供できる点も大きな特長です。

スマートレシートから得られた実際のレシートデータを用いて，開発した自動分類技術の分類精度を検証した結果，インストア商品に手作業で分類名を付与した場合に対し，開発した技術を用いた場合は87％以上の正解率で分類できました。

今後の展開

今後は，分類精度を更に向上させるため，対象カテゴリーの細分化や，量目・調理法情報の抽出を行うことを進めていきます。また，レシートデータに特有の表記揺れや略語といった問題への対応技術は，外食産業への応用も進めていきます。これら以外にも，この技術を購買データ分析基盤に組み込むことで，マーケティング施策の高度化や商品開発支援など，様々なサービスへの展開が可能です。これにより，消費者行動へのより深い理解と，企業の競争力強化に貢献していきます。

文　献

(1) 東芝データ(株)．“レシート印字名に基づきJANコードがない商品をAIで自動分類する技術を開発 ～商品の分類作業を省人化し，レシートデータの分析による商品開発やマーケティング活動の活性化に貢献～”．ニュース．<https://www.global.toshiba/jp/news/data-corp/2024/11/20241107.html>，（参照 2026-01-05）．

真矢　滋
総合研究所 AIデジタルR&Dセンター システムAI研究部