AIを活かすデータクレンジング

McKEN McKen Career Consultants Ltd.

「AIは嘘をつく」

最近こういう話を時々耳にしませんか？

AIはビジネスだけでなく日常生活にも欠かせないものになっていますが、明らかに間違っている情報を平気で提示して来ます。

分かり易い例に、ショッピングサイトでよく目にする次のボタンがあります。



このボタンは商品検索結果一覧の中に表示され、同じ商品なら安く買いたいというユーザー心理に応えてくれる代表的なAI機能です。が、押して表示されるリストの中に直前の一覧で見た安値商品が含まれていないことがよくあります。

裏で動いているAIは商品の様々な属性を解析して同じものをグループ化し安値比較を行いますが、多くの出店者がバラバラに登録する商品情報を正しく判別するには限界があったり、逆に少しの表現の違いから別物と判定したりします。

「AIの嘘（ハルシネーション）」に気づいたユーザーは検索画面に戻ってやり直したりしますが、そのまま少し損をしているユーザーも多いと思います。

AIは膨大な情報を学習して答えを導き出しますが、学習の道筋を間違えれば誤った答に辿り着きます。そこで重要になるのがAIの学習を正しく導く**「教師データ」**であり、それを整えるのが**「データクレンジング」**の役割になります。

「教師データ」と「データクレンジング」

McKEN McKen Career Consultants Ltd.

「教師データ」

こちらも、日常的に使われるレシピサイトを具体例に挙げて説明します。

検索ワード	掲載点数
アボカド	55,781
アボガド	55,781

「アボガド」と間違えて検索しても「アボカド」と同じレシピが表示されます。『「アボガド」と「アボカド」は同じものだと検索エンジンに認識させる』これが「教師データ」の役割です。

では、地方で呼び名が異なったりする食材ではどうでしょうか？

検索ワード	掲載点数
ひれ	599
ヒレ	599
ヘレ	599
フィレ	3,469

関西で「ヘレ」と呼ばれる「ヒレ肉」は関西人の入力がしっかりカバーされていますが、洒落た人が使いそうな「フィレ」になると魚が加わってしまい検索結果が大きく異なってしまいます。

そこで、肉料理に絞りたいので「肉」を検索ワードに加えると…

検索ワード	掲載点数
ひれ肉	896
ヒレ肉	896
ヘレ肉	43
フィレ肉	63

このサイトは、ひらがなとカタカナの違いはシステム的に上手く処理しているようですが、しっかりとした「教師データ」を整備していないので、Filet部位の肉を使った料理を考えている関西人や洒落た人には十分にアピール出来ていないようです。

「データクレンジング」

データクレンジングの基本は曖昧な語やデータのズレを「名寄せ」して「同じもの」と認識することですが、上記のような場合に、様々なユーザーが色んなワードで検索しても同じレシピを提示できるよう検索エンジンに指示する「教師データ」を整備することも「データクレンジング」であり、前頁のAIを活かすことにも繋がります。

今までに対応したデータクレンジングの事例

McKEN McKen Career Consultants Ltd.

1. 購買データ（研究用消耗品）の分析とクレンジング

外資系製薬会社にて20年以上放置されていた試薬購買マスタに基づく購買実績3年分を分析し、購買額の8割を占める8万品目をクレンジングし約2万品目に集約

2. 試薬管理システムと購買システムの連携

国内大手製薬会社の試薬購買実績を分析整理し、新たに導入した試薬管理システムと両システムのマスタを紐付けてシステム連携を実現

3. 間接材購買の商流整理をサポート

複数の国内大手製薬会社にて間接材購買実績を分析整理し、購買金額および購買頻度が高いメーカーの代表的な購買品をリストアップして見積用資料を作成

各社の購買部門になり代わって見積引合いを行い、商流の整理をサポート

- 対応規模：　　数千件～数万件
- コストダウン効果：　　10%程度（初回）

多くのデータを扱うことにより「データを見る眼」が鍛えられました

データの「オカシイ」の本質を見抜く

McKEN McKen Career Consultants Ltd.

「データを見る眼」について最近遭遇した事例を一つ紹介します。

DXを積極的に推進している企業の工場を見学させて頂き、管理の難しい商品を現場の知恵を活かしてうまく扱っておられるのを拝見しました。商品の特性上個体管理は行えないので、商品毎ステータス毎に数量で管理しておられました。

業務に合わせてカスタマイズされた基幹システムは、よく考えられたコード体系の下夫々の商品が管理され、「キー」になるデータに揺らぎを生じさせない仕組みになっているため、「データクレンジング」のニーズがないことはすぐに分かりました。

一方、商品毎ステータス毎に管理されている在庫表の数字に違和感を感じました。

工場では、一つの作業工程を経るとその商品のステータスが変わるので、在庫表の数字は常に動いており、システム上で確認できる商品毎ステータス毎の在庫はその一瞬を切り取ったものにしか過ぎないということで重視されず、数字がおかしくても勘と経験でカバーして工場内の各作業は問題なく行われています。

基幹システムを改めて見させて頂いて気づいたのが「各工程がシステム上前工程を引継がずに処理できる」仕組みになっており、在庫表上で数字が繋がらない原因になっていることでした。（お客様には報告済み）

工場では節目毎に人手をかけて棚卸しを行い、生産計画等はこのリアルな数字を使って定期的に見直されるので、今のままで問題なく業務は遂行されます。

ただし、会社が推進しているDXで上流工程のデータが整えられても、データが内部で繋がらない基幹システムで有効に処理することは難しいのではないかと思われます。