

画像内容に基づく検索技術に対する期待と現実

国立情報学研究所 北本朝展

画像検索とは、画像データベースの中から利用者が求める画像データに素早く到達するための技術を指す。画像検索はいくつかの技術要素から構成される。まず入力方法としてキーワード（タグ）、テキスト、画像などを選び、検索技術としてインデックス、特徴表現、類似度（距離）定義、フィードバックなどの要素に技術的な工夫を加え、検索結果として画像やテキストなどを出力する。以上の技術的要素を様々に組み合わせて用途に合わせた画像検索を実現することになるが、中でも以下の 5 つの組み合わせがよく用いられる。(1) テキストを入力し、画像に付与された統制語やタグなどを検索する。(2) テキストを入力し、画像に関連付けられたテキストを検索する。(3) 画像を入力し、画像から抽出したエンティティを検索する。(4) 画像を入力し、類似した画像を検索する。(5) テキストを入力し、画像の自動アノテーションで得られたテキストを検索する。従来型のデジタルアーカイブでは(1)の方式が多く、Google 画像検索のようなウェブ画像検索では(2)の方式が多い。これらも画像検索の一種ではあるが、あくまで画像と関連付けられたテキストの検索であって、画像内容が主な分析対象ではない点に注意しておきたい。

これに対して画像内容に基づく検索、すなわち画像に付与された付加情報ではなく画像に含まれる内容そのものの分析に基づく検索手法にも長い研究の歴史がある。この検索手法は主に 3 種類に分類できる。(1) 同じもの（エンティティ）を探す。(2) 同じ種類のもの（カテゴリ）を探す、(3) 似ているもの（距離、類似度）を探す。(1)の方法は顔認識を用いた人物検索などが代表的な例であるが、物販（ファッション等）や広告などの分野を中心に商品やブランドの検索などへの応用も進められている。(2)の方法は画像への自動タグ付与に相当するものであり、写真アルバムやストックフォトの整理に使えることが期待されている。一方(3)の方法は「あんな画像を検索したい」という漠然としたニーズに対応できることが期待されているものの、ニーズが漠然としていることもあって、それほど目覚ましい進展はないというのが現状である。

画像検索の研究はこれまでも地道に積み重ねられてきたものの、その進展はテキスト検索などに比べるとゆっくりとしたものであった。その流れを一変させたのが深層学習（ディープラーニング）の登場である。これはニューラルネットワークを用いた機械学習の一種であり、その原理は 1980 年代から知られていたものの、ここ 5 年ほどの間に技術が急速に発展したことで、画像検索の分野にも非連続的な進化が生じることになった。例えば画像の自動タグ付けは人間の精度に匹敵するほど向上し、1000 種類以上のタグを自動付与することさえ可能となった。また画像の自動キャプション付与では、画像を与えるだけでそれを説明する文章が生成できるようになった。このように、深層学習の登場以前はしばらく先の技術とされていた自動生成タグやテキストを用いた画像検索についても、今や実用化へ向けての期待が急速に高まりつつある。

とはいえ、これらの技術が現実的にどれほど有用かは冷静に見極める必要がある。深層学習で大きな成功を収めているのは「教師あり学習」、つまり教師データに合うような出力を生成できるアルゴリズムである。このときアルゴリズムが「正解」を出力できる理由は、概念を理解できたからではなく、関連するデータをたくさん学習したからである。逆に言えば、深層学習モデルに概念を明示的に教えることはできないため、大量の良質な教師データを確保することが精度向上には不可欠となる。こうした良質のデータは、Google や Facebook などの巨大企業であれば自社サービス経由で収集できるが、一般のプレイヤーがそれに匹敵するデータを確保することは至難の業である。そこで、クラウドソーシングやヒューマンコンピューテーションなど、インターネット上の不特定多数の力を集約して生み出された大規模な教師データを研究者コミュニティで共有して利用し、それを共通の基準として研究を比較するアプローチが広がりつつある。

ただし人文学において期待される専門的な情報の検索は、大規模教師データを深層学習に適用するというアプローチとは相性が悪い面もあるため、このアプローチが専門性の高い分野でも有用かどうかは検証する必要がある。第一に、専門的な情報を正確にタグ付けするには専門知識が必要となるため、インターネット上の不特定多数を対象としてさえ適任者の絶対数が少ないという問題がある。第二に、そもそも適任者が少ないため収集できる教師データのサイズも小規模となり、教師あり学習の精度が上がりづらくなる問題がある。第三に、古典籍に現れる線描の絵画のように抽象度が高いデータでは、単純に見えるデータに高度な意味が埋め込まれているため、概念を理解しない機械学習では少数の教師データを有効に活用できない可能性がある。

このように概念という「意味」の扱いは依然として難題であるため、画像検索の研究戦略においては「意味」という問題をできるだけ避けて、物理や統計などの扱いやすい方向からアプローチすることが望ましい。例えば深層学習では、画像の局所的な特徴から画像の大局的な特徴までを多層で表現するモデルを用いるが、そこでは「意味」はニューラルネットワーク上で分散的に表現されているものとされる。ニューラルネットワークが正解を出力するというゴールに至る中間表現においては、意味を明示的に意識して操作する必要は特にないとも言える。

そうは言っても「意味」を避けるだけではいずれ限界に到達する。その限界を越えるには、人間と機械の協調というアプローチ、すなわち人間と機械がチームを組んでお互いが得意なタスクを担当することで、チームとしてのパフォーマンスを最大化することが必要となる。その過程では、大量・高精度な作業を高速化・自動化するタスクを中心に、一部のタスクで人間から機械への置換が進んでいくだろう。しかし高度な判断を下す部分については依然として人間に優位性があるため、少数かつ高度な判断を要する部分では人間を支援して効率化することが機械の役割となる。こうした人間と機械のチーム内分担をどうデザインするかは今後の重要な研究課題である。そして、デジタル人文学でも大規模画像データベースの重要性が増しつつある現在、画像検索技術の継続的な改善には情報学と人文学の研究者が協力するチーム作りも重要な課題となる。