

グロッサリ

非構造化データ

文書・画像・音声・動画などのデータを指す。従来のリレーショナルデータベースでは、最初にデータ構造を定義するが、そのような処理には適合しづらいタイプのデータである。

準構造化データ

センサで取得した数値データや、文章やログなどのテキストデータを指す。データ収集時点では構造化されていないが、非構造化データに比べて比較的構造化が容易であることが特徴である。半構造化データとも呼ぶ。

ライフログ

人間の行動をデジタルデータとして記録すること。個人が意識的に記録するケースと、インターネットアクセスや電子機器の利用履歴を自動的に記録するケースがある。

センチメント分析

文章内のキーワードを自動的に分析することにより、書き手の感情（好意的なのか否定的なのか）を分析すること。

機械学習

人間が過去の経験から新しい知見を得るのと同様に、コンピュータが過去のデータから規則や判断基準などを抽出して、状況の同定や将来予測を可能にする技術。

MapReduce

多数のコンピュータを用いて大規模データを並列処理するためのフレームワーク。データを細かい単位に分割して各ノードに配置する map 処理とそれぞれの処理結果を集約する reduce 処理からなる。

Hadoop

大規模データの分散処理を支える Java ソフトウェアフレームワークであり、Apache ソフトウェア財団からオープンソースソフトウェアとして配布されている。

Cassandra

Apache ソフトウェア財団から提供されているオープンソースソフトウェアの一つで、キーバリュー型分散データベース管理システムである。

Flume

分散ノードから大量のログデータを効率的に収集、集約、移動することを目的に開発された分散型のオープンソースソフトウェア。

(赤津雅晴)