テキストマイニングとは

【 ぱっとマイニングJP 】っていったい何？ —— 形態素解析によるデータの定量化

- テキストマイニングとは、さまざまな文書（テキストデータ）の中から有益な知識や情報を取り出そうとする技術です。
- インターネットで集めたアンケートの結果や、お客様センターにかかってきた問合せの内容、営業マンの報告書、専門分野での論文の束など、デジタル化された文書が大量にあるとき、それを効率よく選別し、いくつかの言葉をキーにして検索したり、全体の傾向を読み取ったり、ほしい情報を抜き出したり、抜き出した情報を分析して理解したり、分かりやすいようにグラフや図に置き換えたりすることがテキストマイニングです。
- かつては手書きされていた文書類が、今ではデジタルデータで保存される場合が多くなってきました。そのため、この【ぱっとマイニングJP】のようなコンピュータ・ソフトを使ってテキストマイニングする機会が増え、マイニング技術の必要性がどんどん高まってきています。
- ○×の数や数字の羅列といったようなデータなら、コンピュータで解析するのはカンタンです。たとえば野球選手の成績のように、あらかじめ蓄積されたデータがあれば、年間の平均打率だったり、左投手に強い・弱いなどという分析がすぐにできてしまいます。ところが、文章のように数字ではない文字列を解析し、数字に置き換えたりするのはカンタンなことではありません。
- しかし、情報科学・学問が発達して文章を定量化（数に置き換えたりすること）することができるようになり、コンピュータ・ソフトによる解析が可能となりました。
- テキストデータの定量化の手法のひとつとして『形態素解析』があります。形態素とは「意味を持つ最小の言語単位」という意味です。「われわれはロボットだ」という文章があるとき、「われわれ・は・ロボット・だ」と、まるでロボットがしゃべるように区切るそのひとつひとつの単語が形態素になります。
- 文章を形態素単位に分解し、形態素の出現頻度を見たり、どの形態素のとなりにどの形態素がよく出てくるかなどを見れば、ひとつひとつの文章を詳細に読むことなく、文章の全体像や傾向などがわかります。
- さらに、結果を見やすい形に変えてみると、文書全体の傾向や特徴が一目でわかるようになり、その上、単に文書を読んでいるだけでは気づかなかったような新しい事実を発見することができる場合もあります。