

■小山照夫 情報社会相関研究系 教授

【タイトル】

より有効な情報検索システムの構築を目指す

【本文】

情報検索システムは、今日なくてはならないものになりました。パソコン、インターネットなどにある膨大な電子データから望む情報を取り出そうとすると、私たちはよく、キーワード検索を行います。しかし例えばインターネット上に現在普及している検索エンジンでは、必ずしも十分に満足する検索結果を得られるわけではありません。それはなぜなのか、そしてどうすれば膨大な情報のなかから必要な情報を選び出すことができるのか。それが私の研究の大きな目標です。

もちろん、一足飛びにこの問題が解決できるわけではありません。とくに、私たちが日ごろ使っている「自然言語」は、表現力が豊かであるがゆえに多様な意味をもち、そのため目的とする情報が取り出しにくくなっています。

まずは専門的な文書から

私がいま研究を進めているテーマの1つは、日本語学術論文の用語抽出です。専門性が高い文書は、含まれる用語の意味が比較的統一されている傾向があります。また、自然言語では「認める」「表す」といった“動詞”を中心に記述が行われますが、日本語の学術論文では「認識する」「表現する」といった“サ変名詞”が多用されるため、用語抽出がしやすいのです。

そこで、情報処理学会の抄録約2万8000件から、サ変名詞を中心とした“動詞概念”と、「データベース」「アルゴリズム」といった“名詞概念”を抽出しました。そして、「利用」など一般的すぎる用語を外したうえで、それぞれの専門用語の関連性、つまり文章内の近い位置に表れる頻度を計って用語を分類しました。すると、分類した用語群から逆に、特定の研究領域を推定できることがわかりました。つまり、キーワード検索をするとき、目的分野の情報を選択的に抽出することが可能になるのです。

それ以外にも、複合語内の関連性についての研究も行っています。たとえば、3つの要素からなる「個人情報環境」という用語が抽出されたとき、これが「(個人-情報)-環境」か「個人-(情報-環境)」かで、意味が違ってきます。前者は情報の内容は個人についてであることがわかりますが、後者では情報内容が特定されないからです。

より一般的な文書へ

先ほど例にあげた情報処理学会の抄録は、1件が300字以内と短い文書です。今後は、論文そのもののような長い文書を扱っていきたいと思います。論文は記述形式が決まってい

るので、セクションごとに分けて 1000 字規模にすることで対応できるでしょう。また、「辞書」「言語」のような専門用語として使う一般語や、言い回しの違いなどについても抽出し、用語分類できるようにしていきたいと考えています。

そうして、専門性の高い文書から徐々に一般的なものへと研究を広げていき、将来はすべての文書についてより有効な情報検索が可能になるようにしていきたいと思います。

(取材・構成 吉戸智明)