

情報学プリンシプル研究系准教授 定兼 邦彦

## 情報社会での平等な「知の共有」のために

### 膨大なデータを処理するために必要な技術とは

現代のような情報社会では、日々増えていく膨大な量のデータをいかに効率良く処理できるかが非常に重要です。具体的には、大量のデータをできるだけ軽量に圧縮しながら、索引部分を高速に検索できる技術を実現する必要があります。ですがこの双方を両立することは、それほど簡単ではありません。

布団を例に考えてみましょう。布団は圧縮袋に入れて余分な空気を吸い出すことで、場所を取らぬようコンパクトに固めることが可能です。とは言え、これは収納をラクにするためであって、布団として活用するには、もう一度空気を入れて元の大きさにふくらませる必要があります。データも同様に、軽量に圧縮したとしても、使用するために再び復元しなくてはならないなら、圧縮した意味がなくなってしまいます。

このような問題を解決するには、必要な部分のデータのみを解凍するという技術が有効です。カップラーメンは、水分を抜くことで「圧縮」された食品ですが、お湯を入れて何分間か待たなくても、箸を入れれば食べたい所がホカホカと瞬時に「解凍」されてつまめる——これが、一部のみを復元して、軽量さと高速な検索とを両立させる技術のイメージに近いかもしれません。このように圧縮率の高さと、検索の高速性のバランスをどのように取るか、ということを念頭に置きながら、日々研究を進めています。

### 玉石混淆の Web から真に役立つ情報を探す

私の研究は、テキストデータのようなシンプルなデータ構造だけでなく、さらに複雑な木構造にも対応しています。木構造は、1つのデータが複数に枝分かれして広がっていく家系図のような形ですが、これがさらに複雑になったものをグラフ構造と呼び、各データが複雑に結びついた網の目のような姿をしています。多くの方が毎日使いこなしている Web がその代表例です。次のステップとしてこの構造にも対応する圧縮、検索の技術を実現することが、私の現在の目標です。

テキストデータでは、設定された索引が精密であればあるほど、単に特定の単語を探せばかりでなく、名詞と動詞が近くにあるケース、といったより高度な検索が行えます。ですが、データ構造自体がさらに複雑なグラフ構造では、「コミュニティの検索」といったことも可能です。リンクの集中度から特定の話題に関する情報量が多い「コミュニティ」の

存在を特定することができるのです。このような場所は、その事柄に関して最も役に立つ、正確な情報が得られる可能性が極めて高いと考えられます。つまり、検索技術に関する知識が全く無い人でも、インターネット上に散らばる知識の宝庫を、縦横無尽に活用できることになる、ということです。やや大げさですが、こうした技術により「知の平等」が実現する、と言っても良いかもしれません。

私のもう一つの夢は、研究成果が一般の方のパソコン上に乗ることです。私の研究は情報社会のデータ流通を支える基盤的な技術に関連しているため、日常生活で一般の方の目に触れることは、ほとんどありません。ですが自分の研究成果が、より多くの方の生活に寄り添い、快適にするお手伝いができたら、これほど嬉しい事はありません。その夢を実現するための第一歩として、文字列を圧縮したまま検索するライブラリなどを無償で公開しています。簡単なプログラミング知識のある方なら利用できるはずですので、興味のある方はダウンロードして下さい。詳しい解説なども、順次、整備していく予定です。

(インタビュー・構成 中野恵美子)