ソーシャルタグを用いたシソーラス構築法の開発

高橋　拓輝

情報検索において、効率的に検索モレやノイズを解消するための手段として、シソーラスを利用しようとする動きがある。シソーラスは人手で作成していたが、近年では Web ページなどを使った自動シソーラス構築が多く研究されるようになった。従来の手法は、上位語・下位語の判定精度が高くない、最新の用語への対応が難しいなどの問題がある。

一方、ソーシャルブックマークサービス（SBS）で用いられるタグは、シソーラス上で扱う語（シソーラス語）との比較研究などから、シソーラス構築に一定の有用性があることがわかっている。タグとして付与された単語対において、出現数にどれくらい差があるか（対称性）と、どれだけ同時に付与されているか（共出現数）の概念を利用すれば、従来の手法の欠点を解決したシソーラスが構築できる、という仮説を立てた。2 つの概念をもとに、上位語・下位語と同義語を判定するための式（BT・ST）を定め、妥当性を検証した。

SBS の 1 つである「はてなブックマーク」から、2000 件の Web ページに付与されていたタグを収集し、単語対の個々の単語の出現数と共出現数を計測したあと、BT・ST を計算した。共出現数と単語の関係性について検討するため、段階的な共出現数の下限を設定し、それぞれの下限以下の単語対を除いた。それぞれの単語対のセットに対し、BT・ST 値それぞれ上位 100 件・下位 100 件にある単語対を既存のシソーラスの単語と比較し、あらかじめ設定したシソーラス語の関係にあてはまるかどうかを調査した。

得られたすべての単語対は 18653 個あり、75%は共出現数が 1 のものであった。また、BT・ST 値それぞれの上位 100 件には、下位 100 件に比べて多くの上位語・下位語もしくは同義語が含まれていた。共出現数の下限を小さくするたびに、上位 100 件中の最低 BT・ST 値が上昇し、上位語・下位語もしくは同義語の取得件数が増加した。縦軸にシソーラス語となる単語対の割合、横軸に最低 BT・ST 値をとってみると、上位語・下位語は 0.4 くらいあたりから漸減する傾向にあるのに対して、同義語は 1.0 の状況をしばらく維持したあと、0.6 ほどに下がるものの、しばらくその値を維持したあと減少するという傾向が示された。

BT・ST 値が高い単語対は、上位語・下位語もしくは同義語になる可能性が高まることが証明され、特に同義語の判定は高い精度を出すことができた。上位語・下位語の判定精度には課題が残った。データのなかには、既存のシソーラスでは定義されていないが、一般には関係があるとされる単語が共出現している例がみられた（「研究」と「研究者」など）。シソーラスの中には、あるキーワードから連想できる単語の一覧を表示した「連想シソーラス」というものがあり、それを構築することへの応用が考えられた。

（指導教員　中山伸一）