

「同時関係分析法」による類義キーワードの抽出（その2）

6P-4

大沢 光 王 晋民

（富士通株式会社・感性技術推進室）

1. はじめに

（1）前報^[4]では、感性工学のための「心理感覚モデル」の構築の方法論である「同時関係分析法」^{[1][2][3]}の「同時関係」の仮説を、①「用語辞典」^[5]の説明文のキーワード、②「学术论文」の著者が付けたキーワード、③「新聞記事」の本文のキーワードに適用して、あるキーワードに関連するキーワードを抽出する原理的な試みを説明した。

この試みでは、1つの条件項と1つの結果項に関する“一重”の「同時関係」を想定したのに対して、本報では、2つの条件項と1つの結果項に関する“二重”の「同時関係」と、条件項と結果項に特定の条件が付いた“条件付き”の「同時関係」などを想定して、関連キーワードの抽出を試みた。また、「同時関係」の確かさによって、関連の深いキーワードを同じまとまりに分類して、対象の「話題」の抽出を試みた。

（2）ちなみに、図1と図2は、“一重”の「同時関係」の概念的な説明図である。①式は、「同時関係」の確かさの値の計算式で、右辺の括弧の中の第1項の前半は、条件*i*の下での条件付きの*j*の出現頻度、後半は、こういった条件がない一般的な*j*の出現頻度であり、条件項*i*から結果項*j*への「同時関係」の確かさの“程度”を表す $[r_{i \rightarrow j}]$ は、この両者の差分である。

図3と②式は、“二重”の「同時関係」の確かさとその値の計算式、図4と③式は、“条件付き”の「同時関係」の確かさとその値の計算式である。なお、実際の「同時関係」の確かさの値の計算では、これらの計算式の統計的な下限値、または、その近似値が用いられる。

$$\textcircled{1} \quad r_{i \rightarrow j} = \max \left(\frac{p_{ij}}{p_i} - p_j, 0 \right) \quad (i \neq j \text{ のとき})$$

$$\textcircled{2} \quad r_{ij \rightarrow k} = \max \left(\frac{p_{ijk}}{p_{ij}} - p_k, 0 \right) \quad (i, j \neq k \text{ のとき})$$

$$\textcircled{3} \quad r_{i \rightarrow j} = \max_{\forall i, j \in S} \left(\frac{p_{ij}}{p_i} - p_j, 0 \right) \quad (i \neq j \text{ のとき})$$

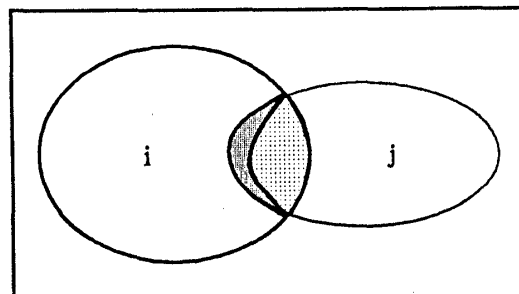


図1 一重の「同時関係」

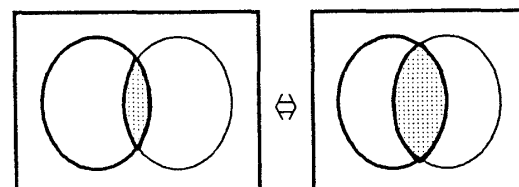


図2 小さな「関係」と大きな「関係」

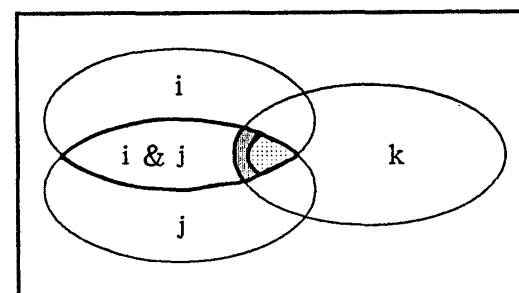


図3 二重の「同時関係」

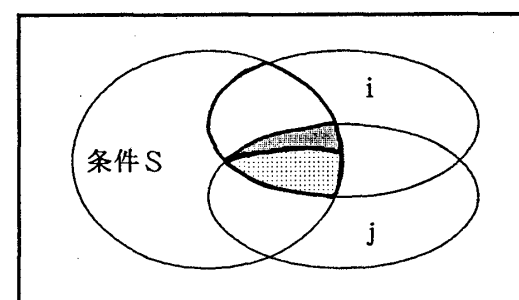


図4 条件付きの「同時関係」

Evaluation the relations between keywords in literatures

by 'simultaneous relationship analysis' (2)

Mitsuru OSAWA and WANG Jinming KANSEI Information Technology Office, FUJITSU LIMITED
H165-1015 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa 211, JAPAN

☎+81-44-754-3277 fax=+81-44-754-3837 NIFTY-Serve=PDG02231 .or. GDD02712

E-mail=osawa@sysrap.cs.fujitsu.co.jp .or. wang@sysrap.cs.fujitsu.co.jp

2. 「同時関係」による関連キーワードの抽出

(1) 『コンピューター用語辞典』^[5]の説明文に含まれているキーワードを例にとり、各見出し語の説明文を1つのまとまりと考えると、“一重”の「同時関係」として、「OS vs オペレーティングシステム」「LAN vs ローカルエリアネットワーク」「CPU vs プロセッサ」「AI vs 人工知能」「DB vs データベース」などの“同意語”の対を見つけることができ、また、「ワークステーション vs パソコン」「UNIX vs 標準」「MS-DOS vs 標準」「ハードウェア vs ソフトウェア」「UNIX vs OS」「ネットワーク vs LAN」「銀行 vs キャッシュカード」などの“内容的に関係の深い”キーワードの対を見つけることができた。

(2) “二重”の「同時関係」では、多くの“同意語”を見つけることができたほか、「パソコン & MS-DOS → OS」「標準 & OS → UNIX」「標準 & 16ビットパソコン → ウィンドウズ3.1」「知識 & 推論 → 人工知能」「プログラム & 言語 → プログラミング言語」など、“一重”の「同時関係」よりも、より“納得できる”キーワードの組み合わせを見つけることができた。

(3) 各見出し語を「ハードウェア」「ソフトウェア」「システム」「通信」「電子デバイス」などのカテゴリーに大別し、「同時関係」の分析の対象を、それぞれのカテゴリーに含まれる見出し語の説明文だけに限定して、“条件付き”の「同時関係」を分析した。結果は、いずれのカテゴリーについても、すべての見出し語の説明文を対象とした“条件なし”の「同時関係」に比べて、“納得できる”組み合わせの割合が多くなった。

ほかの「用語辞典」「学术论文」「新聞記事」についても、それぞれの文献の性格の違いはあるが、ほぼ同様の結果が得られた。

3. キーワードの分類による「話題」の抽出

確かな「同時関係」があるキーワードの対を、その確かさにしたがって、順次、同じまとまりにまとめて、そのまとまりの“意味付け”を試みた。2と同様、『コンピューター用語辞典』を例にとると、“ある”まとまりは、〔OS、UNIX、MS-DOS、標準、16ビットパソコン、ウィンドウズ3.1、マイクロソフト社、OS/2、パソコン、ワークステーション、アプリケーション、大型コンピューター、汎用コンピューター、IBM社、ミニコン、ウィンドウズNT、超大型コンピューター、開発、移植、ソフトウェア、ユーティリティプログラム、コンピューター、RISC、オフコン〕であり、これは、「パソコンOSとパソコン以外のコンピューター」に関する“話題”と考えることができる。また、“別”のまとまりは、〔COBOL言語、プログラミング言語、言語、記述、Fortran言語、プログラム、翻訳、Algol言語、Pascal言語、C言語、アセンブリ言語、外部記憶装置〕であり、これは、「プログラミング言語」に相当する“話題”と考えることができる。

ほかの文献についても、それぞれの文献の性格の違いはあるものが、ほぼ同様の結果が得られた。

【文献】(*は単行本)

- [1]大沢光：「心理感覚モデル」の構築のための「同時関係分析法」、情報処理学会第49回（平成6年後期）全国大会論文集、1995（平成6）年9月
- [2]大沢光：「心理感覚モデル」の構築のための「同時関係分析法」（その2）、情報処理学会第50回（平成7年前期）全国大会論文集、1995（平成7）年3月
- [3]大沢光：「心理感覚モデル」の構築のための「同時関係分析法」（その3）、情報処理学会第52回（平成8年前期）全国大会論文集、1996（平成8）年3月
- [4]大沢光・王晋民：「同時関係分析法」による類義キーワードの抽出、情報処理学会第51回（平成7年後期）全国大会論文集、1995（平成7）年9月
- *[5]大沢光：最新コンピューター用語の意味のわかる辞典（改定3版／第4刷）、日本実業出版社、1995（平成7）年6月

この研究は、通商産業省・工業技術院の産業科学技術研究プロジェクト「人間感覚計測応用技術」の一環として、新エネルギー・産業技術総合開発機構(NEDO)と社団法人人間生活工学研究センター(HQL)を経て、委託を受けて実施したものである。関係の方々に謝意を表する。