連想実験に基づく意味的距離を用いた情報検索*

5L－4

内山 清子†　岡本 潤‡　石崎 俊§
慶応義塾大学大学院 政策・メディア研究科　慶応義塾大学 環境情報学部

1 はじめに

インターネットの普及とともに、全世界の情報を即時に入手することが可能となってきた。一方で、情報が大量にあり過ぎて限定した分野において、本当に欲しい情報をなかなか手にいれることができない。また、キーワードを入力して検索を行なっても、不適切な情報が含まれているため、欲しい情報を入手するために適切なキーワードの設定がわからず、試行錯誤してしまうなど、便利な反面、無駄な時間を費やしてしまうのが現状である。これは、キーワード検索が単なるパターンマッチングによる検索で行なわれているため、情報の内容や意味を理解して検索しているわけではないことが一因である。そこで、本研究では効率的な情報検索のために、情報の内容や文脈に応じた適切なキーワードの選定を目的としている。具体的には、一つの単語を刺激語として、連想する語を記述する連想実験を実施し、連想語を刺激語に対する特徴語とした。その結果に基づいて刺激語で検索した記事に、特徴語が出現する統計を取り、特徴語によって記事ごとの意味的な距離を分析し、情報検索に有効な語について考察する。

2 連想実験による特徴語の抽出

人間の言葉の意味記憶は、辞書に定義してあるような固定したものではなく、さまざまな方向に分散する傾向がある。人間が通常頭の中である言葉を提示した時に思い浮かべる単語は、その言葉となんらかの意味的つながりがあり、意味的な距離が近いものが中心ではないだろうか。そこで連想実験により、刺激語に対して連想単語を調べた。刺激語として、角川書店の類語新辞典を参考にして、自然(気象、天文、気象、植物、動物)、社会、学芸、物品(機械、食べ物、その他)の分野から60語選んだ。被験者には「上位概念」「下位概念」「部分・材料・原料」「属性」「同義語・類義語」「動作概念」「動作環境」という7つの課題を提示して、連想する単語を記入してもらった。一つの刺激語に対して、15から30人の被験者に実験を行なった。連想語はトータルで12728個となった。連想実験のデータを分析すると、馴染みのあるもの、たとえば食べ物や動物などを刺激語にした場合は、多くの連想単語を記述していて、単語の種類がある程度まとまっていることがわかった。また、反対に抽象的な概念(社会、学芸)の分類に属する単語については連想語が少なく、単語の種類も分散していた。課題別の連想語数と平均連想語数を表1に示す。

表1 課題別連想語数と平均連想語数

課題	連想語数	平均連想語数
上位概念	1361	27.20
下位概念	1240	24.80
部分・材料	473	9.46
属性	2028	45.60
類義語	736	14.72
動作	2354	47.80
環境	2036	47.20

実験を行なった刺激語の中から物品の機械に属する単語として「インターネット」を取り上げた。「インターネット」を刺激語とした時に、261個(異なる単語数は55語)が連想語としてあげられた。上位概念、下位概念などの課題を一緒にしてまとめた結果で頻度の高い順に21個を表2に示す。

表2 刺激語「インターネット」に対して連想される高頻度単語(括弧内は頻度)

1	ネットワーク(38)	11	コミュニケーション(5)
2	コンピュータ(31)	11	WWW(5)
3	世界(13)	11	見る(5)
4	電子メール(8)	11	マルチメディア(5)
4	電話回線(8)	11	使う(5)
6	つなぐ(7)	17	サーバ(4)
7	端末(6)	17	楽しい(4)
7	ネットスケープ(6)	17	ネットサーフ(4)
7	ホームページ(6)	17	電子ニュース(4)
7	情報(6)	21	通信(3)
11	ケーブル(5)		

*Information Retrieval using Semantic Distance based on Association Experiment

†kiyoko Uchiyama, Keio University, Graduate School of Media and Governance
‡Jun Okamoto, Keio University,Faculty of Environment Information
§Shun Ishizaki, Keio University, Graduate School of Media and Governance, Faculty of Environment Information

特徴語を設定する際、「インターネット」の連想語として55語抽出されたが、たとえば「つなぐ」「接続」や「楽しい」「面白い」など、意味的に似たような単語があったため、同じ単語としてまとめてた。このようにして30語の特徴語を設定した。

3 文脈における特徴語の出現傾向

今回の連想実験では、課題以外には文脈などの制約が全くないため、刺激語の元来の意味つまりデフォルトの意味についてのデータとして得られると考えた。つまり、30語にまとめた特徴語はテキスト内などの文脈中では独立した特徴語である。テキストでは連想実験による特徴語がどのようになっているのだろうか。そこで、日経3紙のCD-ROM検索システムを使用して、「インターネット」をキーワードにして検索した新聞記事の中からランダムに10文選び出した。その記事の文章に基づいて、抽出された特徴語がテキストの特徴を表すのに有効かどうかを分析した。連想実験により設定した特徴語が記事に出現する回数をまとめた。次に1文書あたりの文字数を一定にして出現頻度を比べるために、10文書の平均文字数を出した。平均文字数が１６２８文字だったので、すべての文書について１６２８文字中に30語の特徴語が何回出現するのかを計算し、平均、標準偏差を計算した。

表3 文書別平均と標準偏差

	平均	標準偏差
文書1	2.15	3.10
2	2.12	3.36
3	3.02	5.57
4	2.92	6.19
5	1.23	1.52
6	1.39	1.86
7	3.33	3.33
8	2.25	2.92
9	1.42	2.87
10	1.85	3.20

選び出した記事の内容はパソコン需要、サイバービジネス、ベンチャー、パソコンネットワークなど様々である。中でも特徴語の平均出現頻度が最も高い文書は文書7で、インターネットの利用者増加に伴うパソコン需要についての記事であり、インターネットの構成、機能、応用などについて詳しく書かれている。特徴語の出現頻度も他の記事に比べるとそれほどばらつきがなかった。「インターネット」という単語そのものの頻度についても9.59と一番頻度が高いわけではなかった。設定した特徴語を基準とすれば、文書7が一番典型的なインターネットに関する一般的な紹介記事である。次に「インターネット」に対して連想される高頻度単語の上位2語「ネットワーク」「コンピュータ」が記事中の頻度の上位2語になっている記事を調べた。表2の上位2語の頻度が高い記事は文書3と文書4であった。「インターネット」の出現頻度は文書3は5.08で文書4は12.53と差があったが、文書3はネットワークの発達について、文書4はインターネットへの接続に関する内容であった。記事の中で出現する特徴語の上位に位置する高頻度単語が、連想実験における高頻度単語のパターンと同じだからといって、記事がインターネットに関する一般的な内容であるとは限らないことがわかった。連想実験による高頻度単語自体はテキスト内容の特徴を抽出するが、頻度パターンについては頻度順ではなく、まんべんなく出現するパターンがインターネットに関する一般的な記事内容であった。

4 おわりに

今回は「インターネット」という単語に注目して、検証を行なったが、多くの単語に有効な、より汎用性のある特徴語をどのようにして抽出していくかを考えていく予定である。また、表2のパターン全体を用いた意味的な距離を評価・分析して、具体的に進めていくことが今後の課題である。

謝辞

研究の協力をしていただいた慶應義塾大学政策・メディア研究科博士課程の今井豊さん及び同大学における連想実験に対する多くの被験者に感謝します。

参考文献

[1] 大熊、石崎、連想実験に基づく概念辞書の構築と検索、情報処理学会秋季大会 (1995)、3H-7、3-47、48、1995

[2] 大熊、石崎、認知実験に基づく概念辞書の構築と検索、情報処理学会自然言語処理研究会報告、NL-112-18、pp.125-132、1996

[3] 大野 晋、浜西 正人、類語新辞典、角川書店 (1981)