技術動向レポート

関連語辞書の自動生成技術の研究開発

―「寄り道検索」が導く新発想―

情報通信研究部
チーフコンサルタント　山泉　貴之

同義語・類義語よりも緩やかな関係を持つ単語の組の集合である関連語辞書の自動生成手法について検討した。関連語辞書によって「寄り道検索」、すなわちユーザが思いつく範囲を超えた情報へ到達するための検索モデルを実現できる。

1. はじめに

スマートフォンまたはパソコン等でのインターネットの利用が日常的となっている現在では(1)、ユーザがスマートフォンまたはパソコン等の検索画面で単語を入力するといくつかの検索結果が表示され、その検索結果の中から知りたい情報にアクセスすることが日常的に行われている。また、それらの検索結果の中には指定した単語と同じまたは類似の意味を持つ単語が含まれることもある。

しかしながら、既存の情報検索システムにおいてはその検索結果が実際に検索を行ったユーザが知りたい情報と異なる場合には、

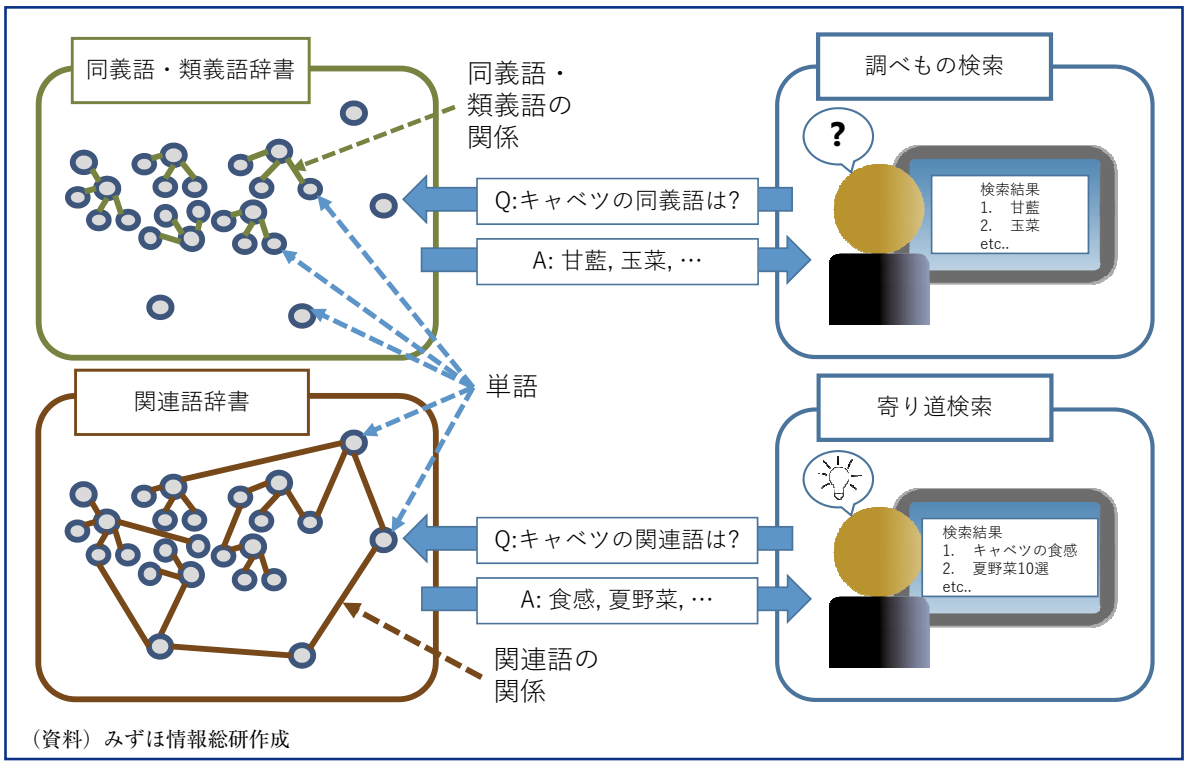
・ユーザ自らが別の単語を指定して検索をやり直す。
・情報検索システムが過去の検索履歴(誤入力やスペリングの間違い)をもとに検索語の候補を提示し、その中からユーザ自らが単語を選択して検索するよう促す。

という手段によって再検索する必要があるため、過去の検索履歴から提示された検索語の候補の中に知りたい情報に関係する単語がない場合や、知りたい情報に関係する単語が思い浮かばない場合には、知りたい情報に到達することが難しいことがある。

そこで、ユーザが真に知りたい情報に到達するまで自力で考えた単語を指定して検索を繰り返し実行する情報検索モデル(以下、「調べもの検索」と記す。)とは異なる新しい情報検索モデルとして、最初に指定した単語をもとにして情報検索エンジンが関連する単語をユーザに代わってその候補を広く提示してユーザに選択を促す等の方法でユーザを新しい情報に到達させる情報検索モデル(以下、「寄り道検索」と記す。)を考える。例えば、「調べもの検索」では最初に「キャベツ」で検索を行っても、「夏野菜についての一般的な情報」を得ることは(「夏野菜」という単語を思いつかない限り、)困難であるが、「寄り道検索」の場合は、「キャベツ」の関連語として「ズッキーニ」→「夏野菜」のように検索語の候補をユーザに代わって考えて提示することで、「夏野菜についての一般的な情報」を得ることができる。これは、ユーザが真に必要な情報が最初に思い付いた「キャベツ」ではなく、「夏野菜についての一般的な情報」であった場合にはユーザの真のニーズを満足させることができるものである(図表1)。

本稿では上記の寄り道検索の実現のために必要な技術として同義語及び類義語よりも緩やか

図表1　「調べもの検索」及び「寄り道検索」における同義語・類義語辞書並びに関連語辞書の利用イメージ



（資料）みずほ情報総研作成

な関係を持つ単語の組の集合である関連語辞書に着目し、寄り道検索に利用できる関連語辞書を低コストで生成する手法について検討する。次に、関連語辞書を利用した寄り道検索の実現の可能性、及び寄り道検索によって開拓可能なマーケットについてもあわせて考察する。

2. 調べもの検索と同義語・類義語辞書

(1)情報検索システム等における同義語・類義語辞書の役割

最初に、関連語辞書よりも厳しい条件の単語の組の集合から構成されると考えられる同義語・類義語辞書について考察する。

ユーザが情報検索システムなどを利用して必要な情報を得る場合、ユーザが真に求める情報へ誘導するための手段として、検索の結果とともに同義語および類義語[(2)]を提示する機能を持つ情報検索システムが日常的に利用されている。同義語や類義語の提示は単語間の関係を定義するためのデータを検索システム等が保持することにより実現される。

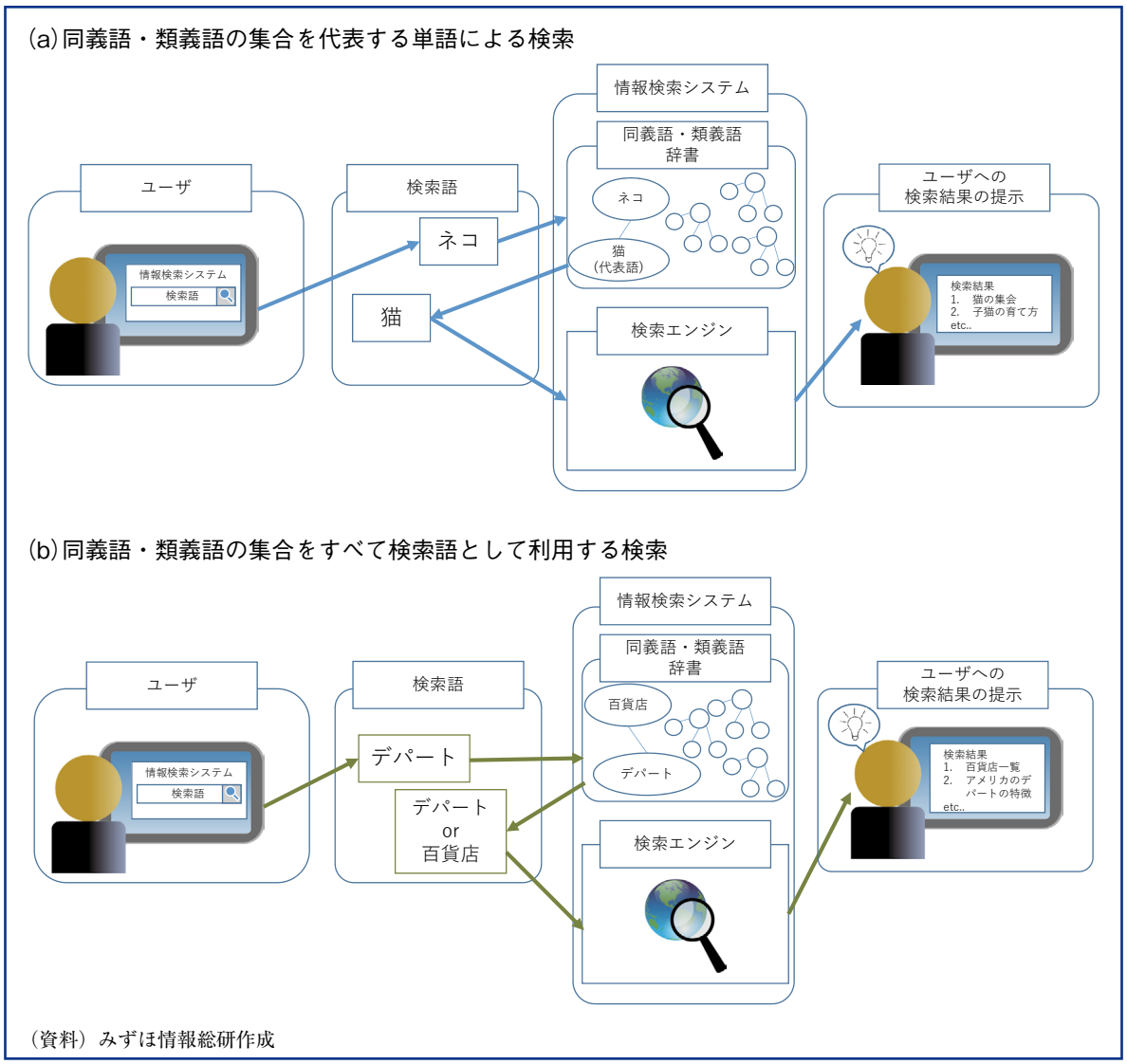
情報検索システム内においては、同義語・類義語辞書は以下の用途に利用されていて、必要な情報への到達性の向上に寄与している（図表2）。

・検索語の送り仮名やかな漢字、外来語などの表記の揺れへの対応（例：「借り入れ」＝「借入」、「猫」＝「ネコ」、「ベネチア」＝「ヴェネチア」）。特に同義語については同義語の集合に対してそれを代表する単語を定義することで、検索エンジン内部における語彙の正規化が可能となる（図表2（a））。

・検索語として指定された語に対する検索結果の他に同義語・類義語の検索結果もまとめて取得して、ユーザに提示する（例：「百貨店」＝「デパート」）（図表2（b））。

・検索語に誤字・脱字が含まれると考えられる

図表2　情報検索システムにおける同義語・類義語辞書の利用イメージ



（資料）みずほ情報総研作成

場合には、単語として正しいと考えられる検索語を推定またはユーザに提示して選択を促すことで、検索語に対応する情報を得る（例：「パンタ」→「パンダ」）。

同義語・類義語辞書を情報検索システムに組み込むことにより、ユーザにとって必要と思われる同義語・類義語を検索結果とともに提示できるため、最短の検索回数で目的の情報に到達することを可能とする検索、すなわち調べもの検索を行う情報検索システムを構築することができる。

また、同義語・類義語辞書は、情報検索システム以外のコンピュータシステムにおける日本語の自然言語処理においても、以下の用途等に利用されている。

・コンピュータ上における文書作成の基本となるかな漢字変換の際の変換候補（入力されたかなに対応する漢字かな混じりの語句等）の提示。

・上記のかな漢字変換の際の変換候補への類義

語の提示[3]。なお、提示される類義語は入力されたかなとは読みが異なっていてもよい（例：「パンタ」→「パンダ」または「パンタグラフ」）。

(2)調べもの検索についての課題

調べもの検索モデルの情報検索システムは日常的に利用されている反面、情報検索システム全体としては以下の課題を内包している。

①検索語の再検討に伴うユーザ体験の低下

情報検索システムを利用するユーザの視点から見た場合、ユーザが最初に指定した検索語そのものが適切でなかった場合には、同義語・類義語辞書による検索語の再選択によって検索を繰り返しても目的の情報にたどり着くことが難しい。適切な検索語を用いた検索が実行されるまでの間、検索語そのものの再検討が繰り返し必要になることで引き起こされるユーザ体験の低下への対処が課題となっている。

②同義語・類義語の判定に伴うコスト

情報検索システムを含むコンピュータシステムにおける上記の用途に使用可能な同義語・類義語辞書を構築するためには、2つの異なる単語を同義語・類義語とすべきか否かの判断が必要となる。しかし、その判断を客観的に行い、かつ完成度の高い辞書を作り上げるためには高度な国語学及び言語学の専門的な知識に基づく判断が必要であることが課題である。

(3)情報検索システムそのものの高度化についての研究の動向

同義語・類義語辞書を用いたコンピュータシステムにここまでに概観したユーザ体験及び辞書の構築時についての潜在的な課題がある一方で、情報検索システムの高度化のための研究[4]等を通して、データを組み合わせて新たな価値を創造することの重要性が指摘されている。もっとも単純かつ直観的なデータの組み合わせの方法として、2つの異なる単語を何らかの基準により結びつける方法が考えられるが、これは辞書を作成する作業に他ならないものである。

3. 寄り道検索と関連語辞書

(1)調べもの検索と同義語・類義語辞書の関係からの類推

調べもの検索の性能向上のためには同義語・類義語辞書が不可欠であるように、寄り道検索の実現のためにも、独自の「辞書、または類似のソフトウェアまたはシステム」の存在が不可欠であると考えられる。

そこで、単語または単語の組について「単語の意味が同一、または似通っているかどうかの判断」を必須としない関連語の組を生成し、その集合体をもって辞書を生成する手法、すなわち関連語辞書を生成する手法について検討する。

関連語辞書は収録の対象となる単語の組を、関連語の抽出の対象となる文書群における単語の組の出現数がある閾値以上であるか否かによってのみ決定する。これにより、「単語の意味が同一、または似通っているかどうかの判断」が必須でなくなるため、その判断のために必要な人的コストを削減することができる。

本稿で検討する関連語辞書を構成する単語の組に類似した概念として、指定された語に対して次の単語の候補としてユーザに提示するための単語の組を集めた辞書があり、はてなキーワード連想語 API[5]や Google サジェスト[6]等がそれぞれ独自に辞書を構築している。これらの辞書は、検索語として得られた単語どうしのみを直接用いて「単語の意味が同一、または似通っているかどうかの判断」を行っていないため、関連語辞書に分類できる。しかし、これらの辞

図表3　同義語・類義語辞書及び関連語辞書と想定される用途の対応

辞書の種類	単語間の意味の同一性または類似性	情報源	構築のために必要なコスト	想定される用途
同義語・類義語辞書	必須	従来型の辞書等	国語学及び言語学の知識が必須	調べもの検索
関連語辞書	必須でない	ユーザが入力した検索語	入力された単語及び付随する情報（指定の順序等）	
	必須でない	一般的な文書等	文書内における出現数の閾値	寄り道検索

（資料）みずほ情報総研作成

書はユーザが検索を行った際に指定した検索語とともに検索語として指定した順序についての情報等を蓄積及び利用することで辞書を構成する単語の組を抽出しているため、ユーザがほぼ同時に思いついた単語どうしが単語の組として抽出されやすい。また、より一般的な文書等から単語の組を自動的に抽出する手法を採っていないことにも留意する必要がある。

ここまでの考察をもとに、同義語・類義語辞書及び関連語辞書と想定される用途の対応を図表3に示す。

(2)寄り道検索の有効性と調べもの検索との関係

同義語・類義語辞書を用いた従来型の検索手法である調べもの検索はユーザの検索についてのニーズが顕在化しているときに有効である。一方、関連語辞書を用いた新しい検索手法である寄り道検索はユーザの潜在的なニーズを顕在化させるのに有効である。

つまり、ユーザのニーズが定まっていない段階で検索する上では、同義語・類義語辞書で範囲を狭めた検索のみで必要な情報を得るよりも、関連語辞書で範囲を拡げて検索を行うことで、ユーザが「実はそれが知りたかった」情報に辿り着き易くする効果を得ることができる。

したがって、寄り道検索及び調べもの検索は互いに排他的なものではなく、ユーザにとってのニーズの顕在度に応じて2つの検索手法を補完的に使い分けることにより、ユーザにとってより最適な検索結果を得ることができると考えられる。

4. 関連語辞書の生成

(1)関連語辞書を生成するための技術的な課題

次に、寄り道検索の実現に必要な関連語辞書の生成手法について検討する。

関連語辞書を生成するためには以下の技術的な課題が存在する。

①関連語辞書に収録するための単語の組を集計するための効率の良いメモリの利用法及び処理方法の確立

関連語辞書は2個の単語の間の関係の強さを評価し、関連性が強いものを抽出することで生成できる。関係の強さを評価するためには、文、文書または文書群（以下、単に「文書等」と書く。）に現れる文字を単語に分解し、その中から任意の2つの単語の組及びその関係の強さを表す量の三つ組のデータを抽出して、その途中経過または結果を配列として保持する必要がある。

2つの単語の組み合わせの数は文書等に現れる単語の数の2乗に比例して増加することから、それらの組み合わせをすべて格納することとした場合、メモリの使用量も単語の数の2乗に比例して増加する。したがって、メモリの使用量を抑制しつつ、2つの単語の関係の強さを評価する手法を確立することを技術的な課題として挙げることができる。

②2つの単語の関係の強さを求めるための評価手法の選択

関連語辞書の生成にあたっては2つの単語の関係の強さ同士を比較する適切な手法を選択する必要がある。2つの単語の関係の強さ及びその比較のための手法としては以下の手法がすでに提案されている。

・共起分析：単語の組が同時に出現したかどうか、また出現した場合にはその出現状況（単語の組の出現回数や連続して出現したか否か等）をもとに単語の組についての関連度の強さなどの分析を行う手法である。

・Word2Vec[(7)], Glove[(8)], fastText[(9)]：ニューラルネットワークを利用して単語に対応するベクトル値を求め、そのベクトル値の近さで関係の強さを求める手法である。

・Poincare Embeddings（ポアンカレ空間への埋め込み）[(10)]：単語をユークリッド空間におけるベクトル値へ変換するかわりに、双曲空間（Hyperbolic Space）におけるベクトル値へ変換する（埋め込み）することでベクトルの次元数を削減する手法である。

関連語辞書を自動的に生成するにあたってはコンピュータのメモリ資源の効率的な利用の観点から2つの単語の関係の強さを表すパラメータをできるだけ少ない次元数で保持することと、関係の強さの計算及び比較についてもできるだけ単純な計算式で実行できることが望ましい。

本稿では、上記の手法のうち2つの単語の関係の強さを表すパラメータ（＝「単語の組の出現回数（または出現率）」）を最も少ない次元数で保持できる共起分析を用いて単語間の関係の強さを求める。2つの単語の関係の強さの比較は単語の組の出現回数（または出現率）を比較するだけで行うことができる。

(2)関連語辞書の自動生成手法とその生成例

関連語辞書生成のためのメモリの効率のよい利用法を探るため、弊社において以下の手順により関連語辞書を自動的に生成する手法を考案し、関連語辞書の生成を試みた。

①大量の文書群（以下、「原文書群」と記す。）から一部の文書群（以下、「サンプリング文書群」と記す。）をランダムに抽出し、共起分析器を用いて共起分析を行う。具体的には、1個の文書内で2つの単語の組（以下、「単語組」と記す。）が出現する文書数を集計し、文書数の分布を求める。

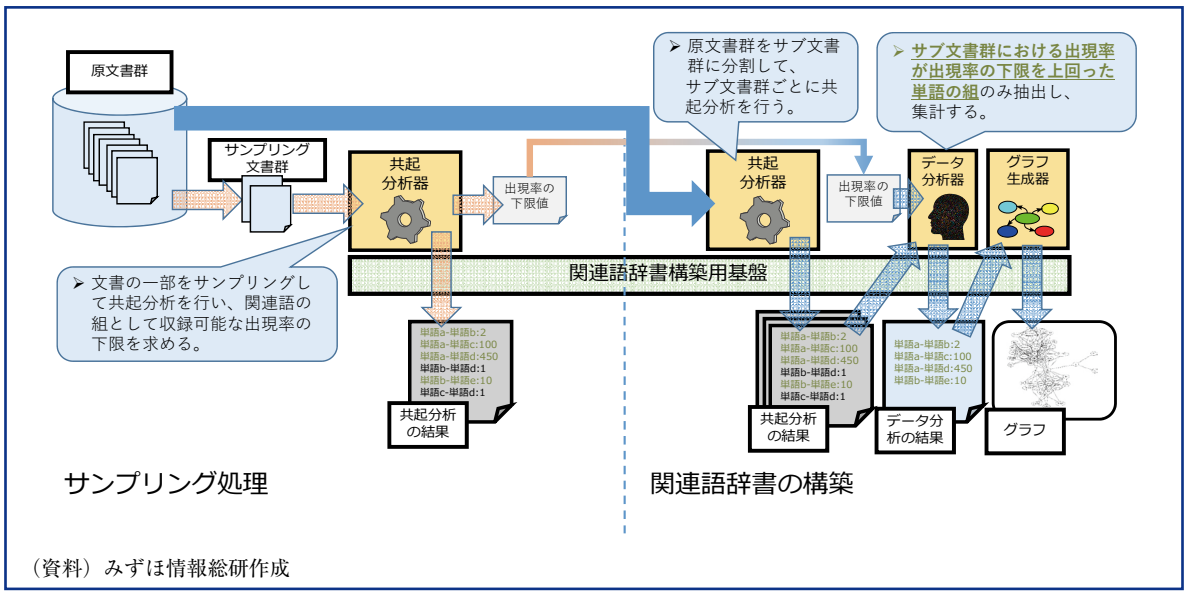
②文書数の分布、原文書群の文書数及びサンプリング文書群に属する文書数（以下、「サンプリング文書数」と記す。）から関連語辞書に収録可能な出現率の下限値[(11)]を決定する。

③原文書群の文書群を複数の文書群（以下、「サブ文書群」と記す。）に分割する。なお、サブ文書群の個数は原文書群の文書数をサンプリング文書数で割った値をもとに決定する。

④手順③で作成したサブ文書群ごとに手順①と同様の方法で単語組のサブ文書群における出現率を求め、データ分析器を用いて手順②で決定した関連語辞書に収録可能な出現率の下限値以上の単語の組のみを抽出する。

⑤手順④でサブ文書群ごとに抽出した単語組の出現数を集計し、関連語辞書を得る。関連語辞書は無向グラフの形式で表現できるので、グラフ生成器を用いてその内容を確認できる。

図表4　関連語辞書の生成例



（資料）みずほ情報総研作成

図表4は上記の手順①～⑤の手法を図解したものである。

5. 関連語辞書の生成結果例と応用例

4節で示した手法により関連語辞書を生成すると、辞書内における単語間の関係性は無向グラフや隣接行列[(12)]等で表現できる。日本語版Wikipedia[(13)]を原文書群として用いて生成した関連語辞書内において、関連性が特に強いと判定できる単語間の関係について無向グラフを用いて描いたものを図表5に示す。

次に、原文書群自体を交換することで専門性を持った関連語辞書が作成できるかどうかについての初歩的な検討を行うために、日本語版Wikipediaから「野菜」のカテゴリに属する文書を抽出して原文書群とし、4節で記述した手法で3,251個の単語及びそれらの単語間の関係から構成される関連語辞書を生成した。さらに、生成した関連語辞書を用いて、関連語と判定された単語をどのように辿ることができるかを確認するため、最初の検索語として「キャベツ」を指定した場合に辿ることのできる単語の例を有向グラフで描いたものを図表6に示す。図表6より、「キャベツ」を起点に「ズッキーニ」→「夏野菜」→「ナス」→「作品」という経路や「食感」→「特産」→「文化」→「キャラクタ」という経路での関連語の検索ができることが確認できる。

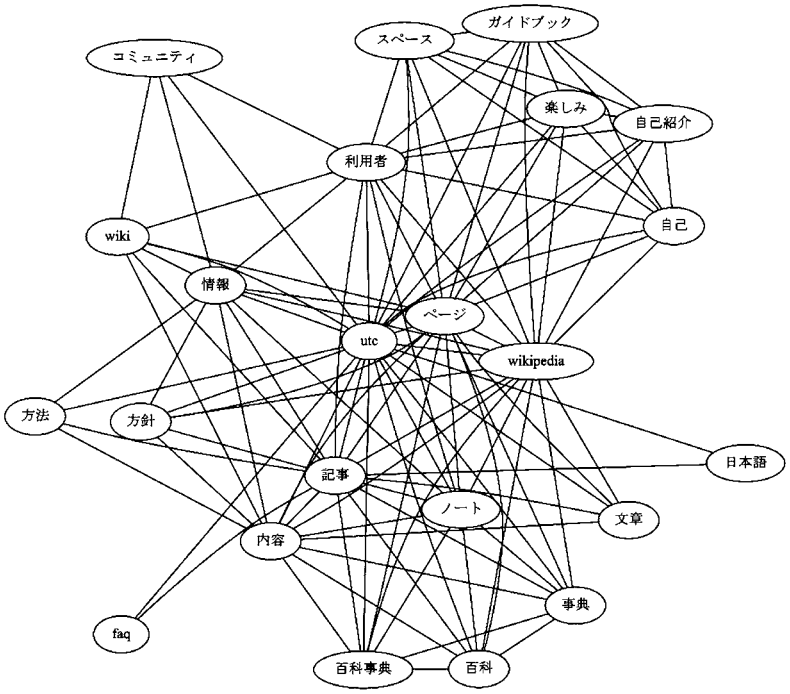
6. 今後の展望及び課題

本稿で検討した関連語辞書の生成手法によって生成した関連語辞書は、文書群の種類を変えることにより、元の文書群が持っていた書き癖や専門性等を反映させることが可能である。また、3節で検討した通り、従来から利用されている同義語・類義語辞書及び調べもの検索との併用も可能である。

関連語辞書は書き癖や専門性等を反映させることができるという特徴を持つことから、以下のような用途例が考えられる。

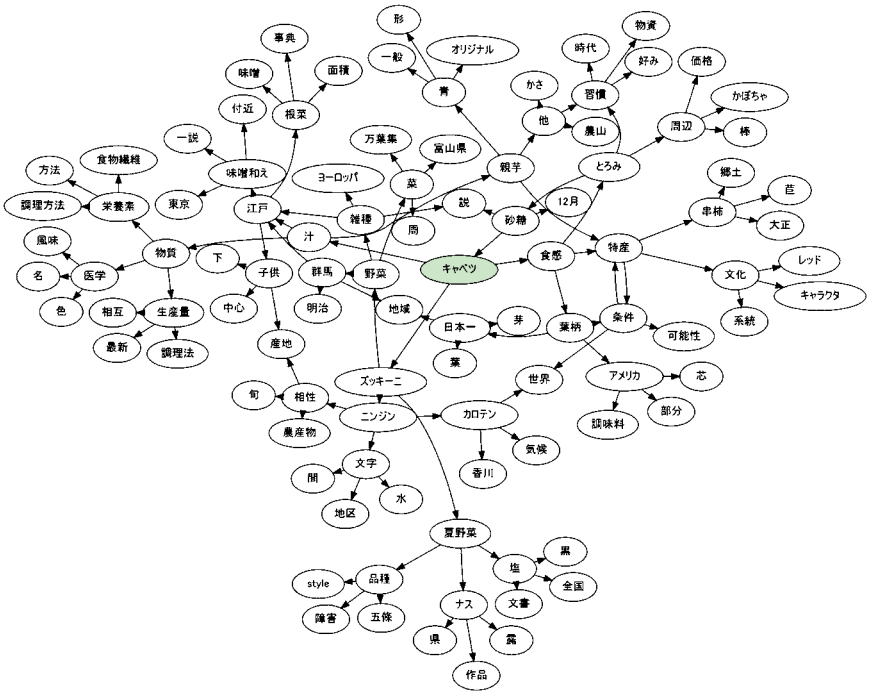
・企業内のドキュメントをコーパスとして用いてその企業が持っている専門性を反映した関連語辞書を生成し、それを利用した関連語の提示機能を情報検索システムに付加すること

図表5　日本語版 Wikipedia を原文書群として用いた場合の関連語辞書の生成結果の抽出例



（資料）みずほ情報総研作成

図表6　関連語辞書により辿ることのできる関連語の経路例



（資料）みずほ情報総研作成

で、企業内の異なる部門間での意思疎通の円滑化を促進する。

・E-コマースのサービスを提供しているWebサイトでは、情報検索システムに関連辞書を用いた関連語の提示機能を追加し、ユーザを寄り道検索へ誘導することで、Webサイトへの滞在時間を増大させ、サイト内の商品にできるだけ多くアクセスさせることにより、ユーザ体験の向上が期待できる。

また、本稿における考察及び検討の結果より、今後は以下の課題についての検討が必要であろう。

・企業の内部に蓄積されている文書群を入力として関連語辞書を生成する場合、本稿において関連語辞書の生成に用いた日本語版Wikipediaと比較して、文書の長さのばらつきが大きいことが考えられる。4.(2)節で試みた手法では、メモリの使用量は文書内に現れる単語数の2乗に比例して増加するため、比較的少ない文書量でもメモリの必要量がコンピュータに搭載されているメモリ量を超えてしまい処理が難しくなることがある。そのため、メモリの使用量を抑制する手法の検討が必要である。

・本稿では大量の文書群を対象とした辞書の生成を想定し、文書群の中からサンプリングして抽出したサブ文書群における単語の組の出現率の分布のみを用いて閾値を設定し、それを用いて関連語辞書への収録の可否を決定している。本稿で検討した関連語辞書への収録条件及び生成された関連語辞書は簡易的なものであり、収録条件及び関連語辞書に収録されている単語の組の範囲の妥当性については詳細な検討が必要である。

注

(1) 「令和元年版情報通信白書」によると、2018年のスマートフォンの世帯保有率は約8割(79.2%)であり、20代以下では(2015年時点における)インターネットの利用時間はテレビの視聴時間よりも多くなっている。

(2) 同義語は意味がほぼ同じ言葉を指し、類義語は相互に変更可能で文脈によっては代替(言い換え)が可能である語で、類語ともいう。

(3) 角川類語新辞典 for ATOK:https://www.justsystems.com/jp/products/kadokawa/

(4) コンテクスト検索エンジン:http://krectmt3.sd.tmu.ac.jp/cse.html

(5) はてなキーワード連想語API:http://developer.hatena.ne.jp/ja/documents/keyword/apis/association

(6) Googleサジェスト:https://support.google.com/ime/japanese/answer/166768?hl=ja

(7) Word2Vec: Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed Representations of Words and Phrases and their Compositionality.", In Proceedings of NIPS, 2013.

(8) Glove: J. Pennington, R. Socher and C. D. Manning, "Glove: Global vectors for word representation," In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543, 2014.

(9) fastText: O. Levy and Y. Goldberg,「Neural word embedding as implicit matrix factorization,」In Advances in Neural Information Processing Systems, 27, pp. 2177–2185, 2014.

(10) Poincare Embeddings(ポアンカレ空間への埋め込み):Maximilian Nickel and Douwe Kiela, "Poincaré Embeddings for Learning Hierarchical Representations," In Advances in neural information processing systems, pp. 6338-6347, 2017.

(11) サンプリング文書群に属する文書の数とサブ文書群(後述)に属する文書の数が一致するとは限らないため、単語組の出現回数に代えて出現率を求めている。

(12) グラフ理論において有限グラフを表すために使われる行列である。行列の要素は頂点の対を表し、その値が0でない場合にはその頂点の対の間が辺によって直接接続されていることを示す。

(13) 日本語における大規模文書群で、かつフリーで利用可能なものが少ないため、本稿においては日本語版Wikipediaを利用して関連語辞書を生成し、それを用いた検索モデルについて検討した。