3N-06

概念辞書グラフの経路に基づく語の直上概念決定法

眞田和枝 † 塩井隆円 † 波多野賢治 †
† 同志社大学文化情報学部

1 はじめに

検索エンジンの利用状況を調査した研究 [1] によると，ユーザの多くは検索エンジンによって得られた検索結果のうち上位 5 件程度しか閲覧しないとされており，インターネットから必要な情報を効率的，効果的に取得することが困難になっている．そのため，ユーザが効率的，効果的に情報を取得するための方法が考えられ、その一例が検索結果の組織化に基づいたユーザへの提示である [2]．

組織化の方法の一つにカテゴリ体系による方法があり，これは，意味のあるラベルの集合であるカテゴリ体系に合わせてデータを整理し，論理的で一貫性のある形にする方法である．しかしこの方法は，カテゴリへの割り当てを自動化させることを可能とする反面カテゴリのラベル集合の定義は人手に頼るところが大きく，大量の文書集合の分類には上手く機能しないという問題がある．

この問題を解決する組織化の方法として，多元的クラスタリングに基づく方法がある．多元的クラスタリングとは，ある類似尺度に準じてデータを自動的にグループ化するものである．この方法ではデータのクラスタリングやラベルづけが自動で行えるが，下位階層のクラスタがユーザにとって直観的でないという欠点がある．これは，多元的クラスタリングが複数の共通素性によりデータが分類されるため，それらの類似度によってラベルが決まるが，ラベル間の意味関係が考慮されず，同位概念ではないラベルを持つクラスタが同位のクラスタとして混在するため，クラスタの構造とラベルの概念構造，特に親子関係に齟齬が生じ，ユーザがクラスタを直観的に理解できないことから生じる．このような問題点を解決するために，ラベルの直上概念の一つを共有していれば同位概念であると判断し，クラスタをラベルの概念構造に従うように提示する方法が考えられるが，多くの場合，それぞれのラベルは複数の直上概念を持つため，ラベルに対する直上概念を一意に決定する必要がある．

そこで本稿では，日本語 WordNet [3] の概念構造を用いてラベルが持つ複数の直上概念から最適な直上概念を決定する手法を提案する．

2 先行研究

Web 検索などの場面において，ユーザが与えた一語のクエリに対して上位語と同位語が大量に得られるが，たいていの場合，上位語，同位語らしくない語が含まれる．そのため，上位語らしい語，同位語らしい語をランキングすることで，それらを判定する手法が提案されている [4]．

文献 [4] ではまず，Wikipedia から抽出した語の上下関係をもとに構築した概念辞書に基づいて，あるクエリ語 q の上位語集合 $H_q = \{x \mid x \in h\}$ と，q を含む q の同位語集合 $C_q = \{x \mid x \in c\}$ からなる二部グラフを構築し，語にハブ値とオーソリティ値というスコアを与えれば，二部グラフの q から h にエッジがあればハブ値が高く，ハブ度が高い h から c にエッジがあればが高くオーソリティ値が高くなるようにそれぞれの値を再帰的に定めることで，最終的に収束したオーソリティ値が高い語ほど同位語らしい語，ハブ値が高い語ほど上位語らしい語としている．

3 提案手法

本稿では，概念間の意味関係を説明した概念辞書である日本語 WordNet の概念構造を用いた，ある語 t に対する最適な直上概念決定を以下の手順に従って行う．

1. 図 1 に示すように WordNet の各概念をノード，リンクをエッジとする有向グラフを構築し，ある語 t の上位ノード $h_1, h_2, \ldots, h_n$ を t の直上概念とする．
2. 共通祖先を持つ直上概念は t に適切な直上概念であると仮定し，$h_i (1 \leq i \leq n)$ の ${}_nC_2$ 通りの組合せの最小共通祖先 (Lowest Common Ansestor: LCA) にあたるノードを抽出する．
3. 抽出した各 LCA から t までの最短経路を求め，

Determining Immediate Hypernym for Words Derived from Path Selection on Concept Dictionary Structure
Kazue SANADA†, Takamitu SHIOI†, Kenji HATANO†
†Faculty of Culture and Information Science, Doshisha University, 610-0394, Kyoto, Japan
{bil0207|bil0225}@mail4.doshisha.ac.jp
khatano@mail.doshisha.ac.jp

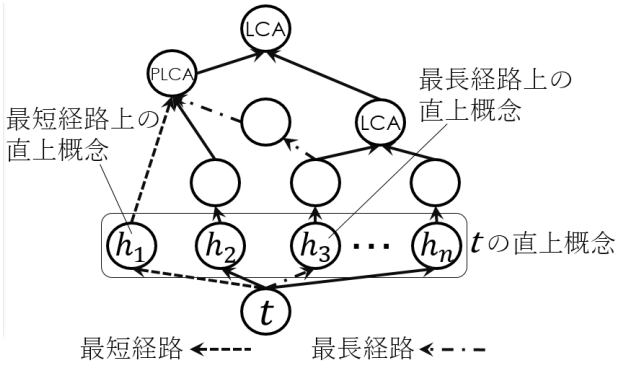

図 1: WordNet のリンク構造を表す有向グラフ

t までの距離が最も短い LCA を最重要共通祖先 (Primary Lowest Common Ansestor: PLCA) と定義する．

4. t と PLCA 間の最短経路と最長経路を探索し，その最短経路上と最長経路上に存在する t の h_i を t の適切な直上概念候補とする．

全てのノードは一つの上位ノードに収束するため，抽出した各 LCA のうち t から各 LCA までの距離が長すぎると LCA が過度に抽象化される可能性がある．それを防ぐために各 LCA から t までの最短経路を求め，t までの距離が最も短い LCA を PLCA としている．また，t から PLCA までの経路上にある直上概念は，t の概念を抽象化し，かつ PLCA を具体化した概念である言える．そのため，t から PLCA までの最短経路は，その概念の階層が最も簡潔であり，その経路上にある直上概念は最も抽象性の高い直上概念と言える．一方，t から PLCA まで最長経路は，その概念の階層が最も詳細であり，その経路上にある直上概念は最も具体性の高い直上概念と言える．

4 評価実験

3 節で定義した直上概念の適切さを判断するために評価実験を行った．

WordNet 内の名詞のうち直上概念を二つ以上持つ概念 t の直上概念を，3 節で説明した「最短経路上にある概念」,「最長経路上にある概念」という二条件によりそれぞれ抽出し,「t の上位概念としてどちらがふさわしいか」を評価した．実験データである WordNet の概念の総数が 1422 であったため信頼率 0.95，最大誤差 0.05 となるように，サンプルサイズを 300 とした．同様に，信頼区間幅の期待値が誤差 $\delta = 3$ 以下となる被験者数もサンプルサイズ $n \geq$20 人となる．δ は予備実験から得た普遍分散を基に求めた値である．

実験の結果から母平均の 95 % 信頼区間を求めたところ，母集団 1,422 概念に対し最適な直上概念は，最短経路による抽出では [816, 891] 個，最長経路による抽出では [559, 634] 個が t に最適な直上概念であると推定され，最短経路による条件が最長経路による条件よりも妥当であると言える．これは，最重要概念ノードまでの最短経路上の概念は祖先ノード数が少なく，より範囲の狭い概念を選択したことから，語 t のみに当てはまるような限定的な概念を選んだために良い結果になったと考えられる．

5 おわりに

本稿では，日本語 WordNet のグラフ構造を用いた，ある概念に対する適切な直上概念決定を行った．評価実験の結果から，ある概念の直上概念うち，PLCA までが最短となる経路上にある直上概念がもっともラベルとして適切であるとわかった．今後の課題として，従来手法と比較し，提案手法の有用性を評価する必要がある．

謝辞

本研究の一部は JSPS 科研費 25540150 の助成を受けたものである．

参考文献

[1] Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Taro Tezuka, Satoshi Oyama, and Katsumi Tanaka. Trustworthiness analysis of web search results. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'07, pp. 38–49, Berlin, Heidelberg, 2007. Springer-Verlag.

[2] Marti A. Hearst. *Search User Interfaces*. Cambridge University Press, September 2009.

[3] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources*, pp. 1–8, 2009.

[4] 佃洸摂, 大島裕明, 田中克己. 上位下位概念辞書を用いた同位語・上位語のランキング手法の提案. In *WebDB Forum 2013*, pp. B4–1, November 2013.