Web 検索を利用した概念ベースの自動構築

奥村 紀之
長野工業高等専門学校 電子情報工学科
noriyuki_okumura@ei.nagano-nct.ac.jp

1　はじめに

本稿では，オートフィードバックを用いた概念の自動学習システムを既存の概念に適用し，概念ベースの自動構築に関する報告を行う．既存の概念ベースに対し，新規の概念を学習させるためにオートフィードバックは良好な結果を示しているが，既存の概念を全て再学習させた場合の検討はなされていない．そこで，概念ベースの評価尺度である X-ABC 評価用データから無作為に抽出した 100 概念に関してオートフィードバックによる学習を行い，関連度評価による検討を行う．

2　関連技術

本節では，実験の基礎となる概念ベース，関連度計算法，オートフィードバックについて解説する．

2.1　概念ベース

概念ベース [1] は，電子化辞書や電子化新聞から自動構築された大規模知識ベースである．概念 A は，その概念から常識的に連想可能な語や，その概念を意味的に特徴付ける語 (a_n) と，その語群を特徴付ける重み (w_n) の対の集合で定義されている (式 1)．

$$A = \{(a_1, w_1), (a_2, w_2), \quad , (a_n, w_n)\} \qquad (1)$$

2.2　関連度計算方式

関連度計算方式は，概念ベースを利用して語と語の関連の強さを定量化する手法である．本稿では，重み比率付き一致度に基づく，重み比率付き関連度計算法を用いて評価実験を行っている．

関連度は 0〜1 の値をとり，1 に近づくほどその概念同士は密接に関係があることを示す．

2.3　概念ベースの評価法

概念ベースの評価は，表 1 に示す X-ABC 評価尺度を利用して行う．この評価尺度に対し，以下の条件を

表 1: X-ABC 評価用データ

X	A	B	C
椅子	腰掛け	机	像
音楽	楽曲	音	電車

満たすとき正答として評価する．MR は関連度の値を示している．

$$MR(X, A) > MR(X, B) \qquad (2)$$

$$MR(X, B) > MR(X, C) \qquad (3)$$

2.4　オートフィードバック

オートフィードバック [2] は，検索エンジンを利用して概念ベースに未定義の語に対する属性を自動収集するシステムである．収集した属性の確からしさは，およそ 70% 程度であることが確認されている．オートフィードバックの流れを図 1 に示す．

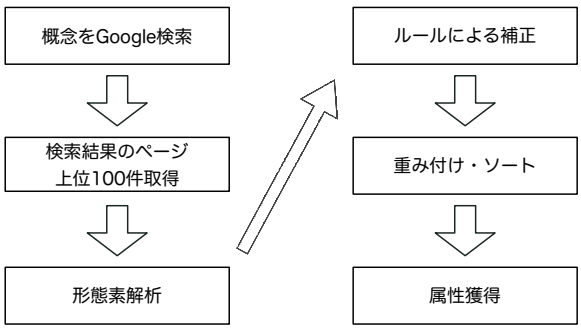

図 1: 属性獲得の流れ

3　評価実験

本稿では，評価実験として表 1 から 25 組，100 の概念をサンプルとして抽出し，複数回のオートフィードバックを実行することにより概念を構築した．

3.1　属性の抽出

先行研究として提案している属性精錬を実装するため，属性取得は複数回行っている．各概念に対し，平均のべ 3000 の属性が取得された．これらの属性に対して，それぞれ概念ベースを再構築し，評価を行う．

3.2　再構築した概念ベースの評価

表 2 に実験結果を示す．オリジナルは既存の概念ベースの評価である．平均は複数回取得した属性の重みを平均した重みを付与したもの，また，精錬は [3] の手法で重みを精錬したものである．

表 2: 評価 (評価用データの抜粋 25 組)

オリジナル	平均	精錬
76%	20%	28%

表 2 に示したとおり，Web より収集した属性を利用して概念ベースを再構築した場合，大幅な精度低下が見られた．しかし，複数回のオートフィードバックを実行し重みを精錬することで，わずかながら精度を向上させることができている．

4　考察

本稿では，電子化辞書や電子化新聞などあらかじめ情報が精錬されている素材から概念ベースを構築するのではなく，Web のような雑多な情報を用いて概念ベースを自動構築することを試みた．

実験結果から，雑多な情報源から概念ベースを構築する場合，単純に属性を収集するだけでは不十分であることが分かった．この原因として，電化新聞などから構築した概念ベースは平均で 30 の属性を保持していることに対し，Web から学習した概念は平均して 3000 もの属性が付与されていることが挙げられる．関連度計算方式の特性上，計算対象となる概念同士で共通する属性が強調されるため，無関連と考えられる概念との間に非常に大きな関連度が算出されてしまうという問題があった．この問題を解消するためには，[4] で提案しているような属性の精錬が必要となる．

5　おわりに

本稿では，概念ベースを電子化辞書や電子化新聞などの整形された素材から構築するのではなく，Web のような雑多な文書から自動的に構築するための問題点について検討した．特に，Web から取得される属性は非常に多く，現在の概念ベースと比較してもその数は膨大であるため，属性の絞り込みが重要であることが分かった．

今後は，統計的に属性を精錬する手法を検討し，整形された素材に頼らない概念ベース構築法を考案する必要がある．また，大島らの手法のように，検索キーワードに対する関連語を的確に取得する手法を組み込む必要がある [5]．

謝辞

本研究の一部は科研費 (23720222) の助成を受けたものである．

参考文献

[1] 「概念間の関連度計算のための大規模概念ベースの構築」奥村 紀之, 土屋 誠司, 渡部 広一, 河岡 司 : 自然言語処理 Volume14 Number5 p.41-64,2007.

[2] 「www を用いた概念ベースにない新概念およびその属性獲得手法」辻泰希, 渡部広一, 河岡司: The 18th Annual Conference of the Japanese Society for Artificial Intelligence, 2004.

[3] 「時間的要因を考慮した属性獲得手法」波多腰優斗, 奥村紀之，情報処理学会第 73 回全国大会，4S-4, 2011

[4] 「Web の時間的変化に着目した未定義概念の属性精錬手法」奥村紀之，言語理解とコミュニケーション研究会（NLC），2011

[5] 「両方向構文パターンを用いた Web 検索エンジンからの高速関連語発見手法」大島裕明，田中克己，日本データベース学会論文誌，Vol.7，No.3, pp.1-6，2008