Wikipediaのリンク構造を利用した商品機能シソーラスの構築

小阪 卓史† 杉本 徹‡
芝浦工業大学 工学部 情報工学科

概要

近年広く行われているネットショッピングでは，Web上の情報をもとに商品を購入するかどうかを判断する．特に，商品の持つ機能は商品の購入を決める際に重要な要素であり，利用者にとって非常に有用な情報であるといえる．

しかし，商品の機能は一般にはなじみのない専門的な用語で説明されていることが多いため，一般利用者が正確に商品知識を得ているとは言い難い．

本研究では，このような専門用語のもたらす情報の欠落の解決法として，専門用語とそれに関係する一般的な語を関連づけた商品機能シソーラスを構築する．

1. 背景と目的

近年，Wikipedia[1]を対象とした研究は広く行われている．中村らの研究[2][3]では記事中のリンクの出現頻度から概念間の関係の強さを算出することで，高い精度で概念の関連語が抽出されている．また，WikiRelate[4]に挙げられるように，記事そのものではなく，記事の属するカテゴリを解析することで概念間の関係を推定する手法もある．

本研究での目標は専門用語についての連想シソーラスの構築であり，専門用語に関係のある一般語を抽出するために，その語がどの程度一般的であるのかを考慮するという点において，先行研究とは異なる．

ここで，専門用語とは「MP3」や「Opera」といった技術的な用語であり，一般語とは「音楽」や「インターネット」など日常において使われる単語のことである．

本研究では，Wikipediaを用いてこのような専門用語と一般語を対応付けたシソーラスを構築する．このシソーラスを商品機能シソーラスと呼び，その妥当性や有用性について検証する．

2．カテゴリ構造を利用した関連概念の抽出

2.1 Wikipediaのカテゴリ構造

カテゴリとは各記事に付与された情報であり，その記事の分野を表す索引の役割がある．また，カテゴリのもつ情報には，そのカテゴリが属するカテゴリである上位カテゴリ，そして，そのカテゴリに属しているカテゴリである下位カテゴリというものがあり，それぞれをカテゴリ間を関係づけるリンクと考えると，カテゴリは巨大なネットワークを形成していると捉えられる．

本研究では，このカテゴリ構造に着目し，概念の一般性を算出する．

Construction of a Product Function Thesaurus Based on the Link Structure of Wikipedia
†Takashi Kosaka
‡Toru Sugimoto
Department of Information Science and Engineering, Shibaura Institute of Technology

2.2 専門用語と一般語

Wikipediaでは，カテゴリには一意性が保証されているため，カテゴリを概念として考える事が出来る．そこで，カテゴリを専門用語の関連語を表す概念として考え，カテゴリの一般度を求めることで概念の一般性とする．概念として記事ではなくカテゴリを選択した理由は，カテゴリが2008年12月時点でおよそ5万件存在していることから一般的な語を網羅していると考えられ，また，記事の50万件と比較して非常に少いため，比較的高負荷な処理も実現可能であるといったことが挙げられる．

2.3 専門用語の関連概念

専門用語の関連概念はWikipediaの記事を解析することで抽出する．本研究では，関連概念を幅広く抽出するため概念間の関係の強さは考慮せず，次のすべての概念を抽出する．

1. 記事中に出現するリンク先記事のカテゴリ
2. 記事のカテゴリ，およびその上位カテゴリ

これにより抽出した専門用語「MP3」の関連概念の一例を示す．

1. 「音声ファイルフォーマット」「デジタル技術」
2. 「コーデック」「データ圧縮」「信号処理」

3．概念の一般性算出

3.1 概念の一般性

本研究では，各概念の一般性はカテゴリの構造に基づいたいくつかの指標を用いて算出する．それぞれの指標，およびその適用理由は次のとおりである．

(1)最上位カテゴリからの距離

カテゴリはその性質上，下位カテゴリよりも上位カテゴリの方がより一般的な概念を表していることから，最上位カテゴリを最も一般的な概念とすると，そこからの距離はカテゴリの一般度に深い関係があると考えられる．

(2)カテゴリに属する下位カテゴリ数

あるカテゴリに属している下位カテゴリの個数は，そのカテゴリが表す概念の一般性に依存することが多い．これは，「音楽」という一般的な言葉に関連する概念は多く，「音声ファイルフォーマット」等の専門用語に関連する概念は比較的少数であることから推測できる．

(3)カテゴリに属する記事数

(2)と同様に，一般的な概念ほどその概念に属する概念は多くなる．

ここで，(1)，(2)，(3)それぞれの指標に関して，あるカテゴリの一般度は，全体のカテゴリの値におけるそのカテゴリの値の出現頻度に依存する．たとえば，(1)の指標は，全カテゴリが0～10までの値をとり，距離5までに約半数のカテゴリが出現する．そこで，距離10のカテゴリを1.0，距離5のカテゴリを0.5として，一般度としている．

また，(2)と(3)に関して，対象カテゴリだけでなく，下位カテゴリにより関連付けれているカテゴリを再帰的に処理を行う．これは，カテゴリとその下位カテゴリ間には一般度にそれほど大きな差が生じないことから，下位カテゴリの一般度もカテゴリの一般度算出に有用であると考えられるためである．そこで，下位カテゴリの一般度の平均を(4)および(5)として，カテゴリの一般度算出の指標として適用する．

以上の指標を用いて，カテゴリ C の一般度を次の式で表現する．

$$\text{Generality} = \sum \text{Gen}_i(C) \cdot W_i \quad (\Sigma W_i = 1.0)$$

ここで，Gen_iは指標(1)～(5)により算出した一般度を表し，W_iはその重みである．

3.2 概念の一般性調査

一般度算出に用いる重みの決定，また，求めた一般度の評価データに用いるため，２０代の学生９名を対象にアンケート形式で１３種類の概念の一般性調査を行った（表２）．一般に，概念の一般性は個人への依存度が高く，回答者の知識量に左右されてしまう．そこで，「その概念が一般の人々の間でどの程度知られているか」という観点に基づいて評価してもらうことで，結果のばらつきの抑制を図った．

表２．概念の一般度調査（一部）

概念	一般度（0.0 ～ 1.0）
音楽	1.000000
IPod	0.694444
インターネット	0.819444

3.3 一般度算出における各指標の重み

各指標の重みは遺伝的アルゴリズムにより決定する．このとき用いる教師信号は，前述の調査を行った概念の中から５つの概念を選択した（表３）．

表３．重み決定に用いた教師信号

概念	一般度（0.0 ～ 1.0）
コンピュータ	0.972222
ブログ	0.666664
コンピュータグラフィクス	0.458333
音声ファイルフォーマット	0.222222
画像ファイルフォーマット	0.138889

これにより，重みは次のように定まった(表４)．

表４．各指標の重み

(1)	(2)	(3)	(4)	(5)
0.153105	0.638244	0.009775	0.188632	0.010244

４．評価

算出した一般度の妥当性の評価は，前述のアンケートにより調査した概念の一般性を用いて行った．教師信号として用いた概念を除いた８概念について，手法により求めた一般度と比較した結果の一部を表５に示す．

表５．算出した一般度とアンケートの結果との比較

概念	本手法	アンケート	誤差
音楽	0.890639	1.000000	0.109361
インターネット	0.864451	0.819444	0.045007
無線	0.803027	0.750000	0.053027
IPod	0.424249	0.694444	0.270195

概念数	一般度に関する誤差					
8	最小	0.04140	最大	0.38954	平均	0.17760

５．商品機能シソーラスの構築

以上をふまえて，専門用語とそれに関係のある概念を対応付けたシソーラスの構築を行った．このとき，一般度について閾値を設けることで一般的な概念のみが関連付けられたシソーラスを構築した．ここで，閾値は 0.7 とした．一部結果を表５に示す．

表６．商品機能シソーラス（一部）

専門用語	関係のある一般概念
MP3	コンピュータ，音楽，ゲーム，社会
Opera	ソフトウェア，インターネット，情報機器
Bluetooth	無線，コミュニケーション，通信プロトコル

６．考察

結果から，本手法により算出した概念の一般性が人の感覚と大きくかけ離れたものでないことが認められた．したがって，その一般性に基づいて構築したシソーラスも十分に妥当であると考えられる．一方で，「MP3」における「社会」のように専門用語にあまり関係のない一般概念が含まれてしまっている．これは，概念間の関係の強さを考慮していないためであり，概念間の関係推定や概念の重要度計算等，一般性とは異なる指標を適用することで対応できると考えられる．

７．まとめと今後の課題

本研究では，Wikipedia のカテゴリ構造に基づいて概念の一般性を算出し，専門用語の関連語のうち，閾値以上の一般度をもつ概念のみを対象とした連想シソーラスを構築した．このシソーラスを用いることで，「音楽」や「インターネット」などの一般的な言葉によってその機能を有する商品を検索できるような，専門知識のない利用者を対象とした新たな商品検索技術への応用が期待できる．

また，調査の結果，算出した一般度が高い精度で人の感覚と適合することから，Wikipedia のカテゴリ構造は概念の一般性算出に有用な情報を含んでいることが認められた．一方で，概念の一般性のみではシソーラスにノイズとなる概念が多数含まれてしまう．今後は，このような問題に対応するため，一般性以外の指標を適用するべきである．

参考文献

[1]Wikipedia : http://ja.wikipedia.org/wiki/

[2]中村浩太郎,原隆浩,西尾章治郎:"Wikipedia マイニングによるシソーラス辞書の構築手法", 情報処理学会論文誌:48,10, pp. 2917-2928(2006)

[3]中村浩太郎,原隆浩,西尾章治郎:"Web 辞典からのシソーラス辞書構築手法", 情報処理学会論文誌: データベース, 48, SIG19(TOD 34), pp. 27-37(2007)

[4]M.Strube and S.Ponzetto: "WikiRelate! Computing semantic relatedness using Wikipedia", Proc. of National Conference on Artificial Intelligence (AAAI-06), Boston, Mass, pp. 1419-1424 (2006)