6C-5

シラバスデータを使った分野ごとの概念マップの生成

下司義寛 † 三輪眞木子 †† 神門典子 ‡ 廣川佐千男 ‡‡
† 九州大学理学部 †† メディア教育開発センター ‡ 国立情報学研究所 ‡‡ 九州大学情報基盤センター

1 はじめに

増え続ける文書を効率的に検索し活用するために、対象とする分野ごとのメタデータや統制語の重要性が高まっている。例えば、複数科目の関連や整合性を調べたり、異なる組織間でのカリキュラムの比較をするためには、それぞれの分野の基本的用語が必要となる。文献 [5, 6] では、カリキュラムの分類や分析のためのシステムが述べられている。情報処理学会においてもカリキュラムの枠組を記述するためのキーワードの検討が行なわれている [3]。

専門用語の構築については、人手によるものや組織的な取り組みの他に、自動的抽出を試みる研究も多数ある。例えば、[4] では比較的長い一つの文章から重要語を抽出している。[7] では、文章そのものの代わりに書籍の索引を利用する方法を提案している。このような専門性の高い少数の書籍を利用する代わりに、短い大量の文書群を対象とする研究 [1, 9] もある。

分野ごとの専門用語のリストそのものよりも用語間の関連がより重要となっている。建築や製造の分野限定のオントロジーの構築事例が報告されている。また、医学・生物学分野を対象としたシソーラス構築プロジェクトも進められている。類似語のように対称性のある関連だけでなく、二つの単語の上位下位関連が多数の専門用語について確定できれば、その分野の構造を表すオントロジーと見なすことができる。

本稿では、Web で公開されるシラバス文書群を対象として、それらに現れる単語の文書頻度を使って専門性の高い用語抽出を行ない、さらに、用語間の上位下位関連の抽出を行なう方式を提案する。

2 文書頻度を用いた概念グラフ構成アルゴリズム

U を文書の全体集合、D をその部分集合、w, v を二つの単語とする。単語 w を含む D 中の文書集合、ならびに w と v の両方を含む文書集合をそれぞれ $D(w)$、$D(w*v)$ と表し、それらに含まれる文書数を $df(w, D)$、$df(w*v, D)$ と表す。ある単語 w を決めたとき、D の範囲で単語 u が w に関連する特徴語であるとは、v を含む文書の過半数に w が現れることと定義する。すなわち、$df(w*v, D)/df(v, D) > 0.5$ となるとき、v は w に関連する特徴語と考える。単語 u と v が $df(u*v, D)/df(v, D) > 0.5$ かつ $df(u, U) > df(v, U)$ という関係を満たすとき、「文書頻度の観点から u は D の範囲で v の上位である」ということにする [2]。概念グラフは、単語をノードとして、上位の単語と下位の単語の組をエッジとする有向グラフである。ただし、上位下位の単語を全て繋いでしまうとエッジが複雑になり全体像が分からなくなるので、ある単語 u よりも上位の単語の中で、上位下位の関係について極小な単語のみについてエッジをつける。ある単語 w の概念グラフは D として U の中で w を含む文書全体としたとき、D の範囲で w に関連する特徴語についての上位下位関連を表すグラフである。

非対称的な単語間の上位下位関連について、[10] では、専門用語を頻度に応じて階層化し、隣接する階層間で類似する二つの単語を繋ぐことによるシソーラス構築法が述べられている。本稿と同様な尺度を用いる研究として [8] があるが、$df(u*v, D)/df(v, D)$ の値を最大とする u, v 間にエッジをつける。これらと本方式との比較は別途発表予定である。

3 実験と評価

本発表で概念グラフの構成に使ったのは国内の大学で公開されているシラバスページ群である。図 1 は、ある大学の「自然言語処理」という科目のシラバスである。これまでの研究で 291 サイトから収集済みの合計 32875 個のシラバスファイルを利用した。

これらのファイル群についてキーワード検索を実現し、与えられたキーワードを含むシラバスとそれらのシラバス群に特徴的な単語を抽出し、さらに単語群の上位下位関連を求めた。例えば、「オートマトン」というキーワードを含むシラバスは 45 件あり、5 個以上のシラバスに出現する特徴語として下の 24 個が抽出された。単語の後の括弧内に数はその単語含むシラバスの個数である。その中の、基礎離散数学、順序集合、情報解析学、小倉久和、離散関係はオートマトンの話しも少しは含まれる広い数理情報の入門的な講義のシラバスにおける関連語であった。実際、基礎離散

Concept Graphs Generated from Univesity Syllabi
† Yoshihiro SHIMOJI(y-shimo@i.kyushu-u.ac.jp)
†† Makiko Miwa
‡ Noriko Kando
‡‡ Sachio HIROKAWA(hirokawa@cc.kyushu-u.ac.jp)
Department of Physics, Kyushu University (†)
National Institute of Media Education (†)
National Insitute of Informatics (‡)
Computing and Communications Center, Kyushu University (‡‡)

科目名	自然言語処理		
開講期	4セメスター	単位数	2
担当者名	松原 康夫	メールアドレス	matubara@shonan.bunkyo.ac.jp
授業概要	私たち人間が使う言葉を、プログラミング言語などの人工言語と対比して自然言語という。自然言語は人工言語よりはるかに複雑なものであり、言語学などの長い研究の歴史がある。その中からChomskyの理論も出てきて、コンピュータ・サイエンスに大きな影響を与えた。また逆に、最近ではコンピュータ・サイエンスを意識した文法理論も出されるようになってきている。この授業では、最初Chomskyの考え方を説明し、次に自然言語を処理する技法の中でも比較的完成されつつある、形態素解析と構文解析を中心に解説する。		
授業計画	1. 言語学から 2. Chomskyの理論 3. 言語理論とオートマトン 4. 形態素解析とその手法 5. 接続表による解析 6. 構文解析アルゴリズムの分類とトップダウン縦形法 7. ボトムアップ横形構文解析アルゴリズム 8. 論理と意味		

図 1: シラバス例（自然言語処理）

数学を含まないシラバスに制限するとそれらは関連語にはならなかった。ちなみに、「ロフト」は「ホップクロフト」の一部と考えられる。

オートマトン (45), 形式言語 (16), 文脈自由文法 (11), 非決定 (8), 帰納的関数 (8), 正規表現 (8), 正規文法 (7), Automata(7), チューリング (7), 正規言語 (6), 基礎離散数学 (6), 決定性 (6), 順序集合 (6), 文脈自由言語 (6), プッシュダウンオートマトン (6), 情報解析学 (6),Ullman(6), 小倉久和 (6), ロフト (6), ホップクロフト (6), 非決定性有限 (6), 離散関係 (6),Chomsky(5), プッシュダウン (5)

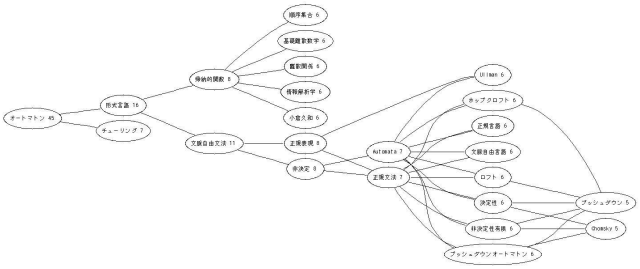

図 2: 「オートマトン」についての概念グラフ

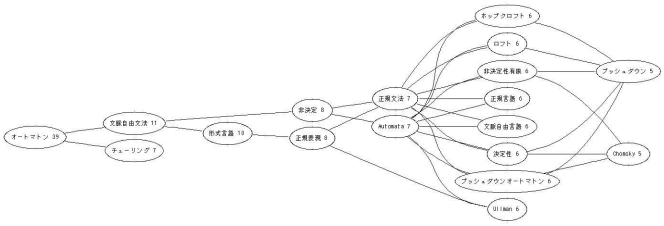

図 3: 「オートマトン – 基礎離散数学」についての概念グラフ

図 2 は、これらの単語の関連の上位下位の関連を表す概念グラフであり、左側が上位、右側が下位となっている。これだけの文書数であっても、上位下位の関連は妥当なものになっていることが図 3 のように確認できた。

参考文献

[1] 藤井, 石川, World Wide Web を用いた事典知識情報の抽出と組織化, 電子情報通信学会論文誌, Vol.J85-D-II, No.2, pp.300-307, 2002

[2] 廣川, 下司, 和多, 文書群からの概念グラフの構成, 情報処理学会第 169 回自然言語処理研究会,pp.79–84, 2005

[3] 情報処理学会アクレディテーション委員会, 情報および情報関連分野における最低水準とは?キーワード案, http://www.myu.ac.jp/ togashi/jabee/FIT2004/pdf/hikita.pdf, 2004

[4] 松尾, 石塚, 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌, Vol. 17, No. 3, pp.217–223, 2002

[5] 宮崎 他, 大学評価・学位授与機構における学位授与のための科目分類支援システムの試作, 情報処理学会論文誌, Vol. 46, No. 3, pp. 782–791, 2005

[6] 野澤 他, シラバスの文書クラスタリングに基づくカリキュラム分析システムの構築, 情報処理学会論文誌. Vol. 46, No. 1, pp. 289–300, 2005

[7] 中西 他, 特定分野を対象とした連想検索のための書籍の索引部を用いたメタデータ空間生成方式, 電子情報通信学会論文誌, Vol.J88-D1, No.4, pp.840–851, 2005

[8] Y. Niwa et al., Topic Graph Generation for Query Navigation, NLPRS'97, pp.95–100, 1997

[9] 桜井, 佐藤, ワールドワイドウェブを利用した用語説明の自動生成, 情報処理学会論文誌, Vol.43, No.5, pp1470–1480, 2002

[10] P. Srinivasan, Thesaurus Construction, in W.B. Frakes and R. Baeza-Yates eds, Information Retrieval: Data Structures and Algorithms, Prentice-Hall, 1992.