# 講義同士の関係構造分析のためのシラバス処理方法の一検討

関谷 貴之† 山口 和紀‡

†東京大学情報基盤センター ‡東京大学総合文化研究科

## 1 はじめに

多くの大学で講義のシラバスを公開しているが，講義内容やカリキュラムを教職員や学生が理解するのは必ずしも容易ではない．そこで我々はシラバスの特徴を分析し，類似関係等を定義して，講義同士の関係構造を全体的に把握する方法，さらにカリキュラム設計に活かす方法を研究している [1]．本報告では，シラバスのテキストに含まれる用語の集合に基づいて，講義同士の関係構造を分析する方法とその分析結果について述べる．

## 2 用語集合に基づくシラバスの構造化

本研究において，シラバスとは講義の内容や授業計画を説明するテキストであり，そのテキストから抽出した講義を特徴付ける概念同士の関係から，講義同士の関係構造を把握するものとする．しかし，シラバスのテキストは分量が限られるため，概念として意味のある単語を抽出するのは容易でない．また多義的な単語があることから，ある概念を一つの単語で表わすことは困難である．そこで，単語の組み合わせとして用語集合という概念を導入する．図1は用語集合の例である．このように，文書(或いは教材) $d_j$ の任意の一つの文に含まれる異なる単語 $w_1, w_2, ...., w_n$ を任意に組み合わせた集合 $\{w_a, w_b, ..\}$ を用語集合 $ts$ と呼ぶ．文書 $d$ は，当該文書に含まれる幾つかの代表的な用語集合 $ts$ によって特徴付けられるものと考える．

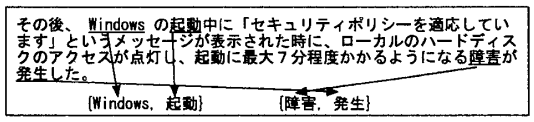


図 1: 用語集合

代表的な用語集合を抽出するにあたり，まず用語集合間の上位・下位関係を定義する．用語集合 $ts_x = \{w_a, w_b, ..\}, ts_y = \{w_p, w_q, ..\}$ について，$ts_x \supset ts_y \wedge |ts_x| = |ts_y| + 1$ が成り立つ時，$ts_x \sqsupset ts_y$ と表記し，$ts_y$ は $ts_x$ の上位の用語集合と呼ぶ．

次に用語集合 $ts$ における tf-idf [2] に相当する値 tsf-idf を式1で定義する．ここで，$tsfn(ts, d)$ は文書 $d$ 中で用語集合 $ts$ を含む文の数，$N_S(d)$ は文書 $d$ 中の文の数，$N_D(ts)$ は用語集合 $ts$ を含む文書の数，$N_D$ は文書の数を表す．

$$\begin{aligned} tsf(ts,d) &= \frac{tsfn(ts,d)}{N_S(d)} \\ idf(ts) &= \log \frac{N_D}{N_D(ts)+1} \\ \text{tsf-idf}(ts,d) &= tsf(ts,d) \cdot idf(ts) \end{aligned} \quad (1)$$

用語集合 $ts$ は単語の組み合わせであるため，構成する単語の個数 $|ts|$ が多い用語集合同士で，同一の単語を複数含む可能性がある．また，専門的な用語では，少数の単語でも意味をなすと考える．そこで，代表的な用語集合 $ts_x$ は，式2を満たすものとする．同一の単語を含む用語集合が複数ある場合，この式では $\delta$ が大きいほど，より少ない単語で構成される用語集合が代表的な用語集合となる．実際のシラバスを用いた予備的な実験より，$1 < \delta < 1.5$ の間で抽出される用語集合の構成が大きく変わることが分かっている．

$$\forall ts_y (ts_x \sqsupset ts_y), \frac{\text{tsf-idf}(ts_x, d)}{\text{tsf-idf}(ts_y, d)} > \delta \quad (2)$$

## 3 シラバスの分析

2節の分析手法を，本学教養学部1,2年生向けの825の講義のシラバスに対して適用した．825個のシラバスから抽出された単語は約10,000個で，用語集合は約7,000,000個である．また，多くのシラバスに含まれる名詞の数は300個以下で，平均で130個程度の単語が抽出された．

単語の抽出には形態素解析ツールMeCab[1]を用いた．また，個々の講義のシラバスから抽出する用語集合の数 $N$ は，シラバスに含まれる単語の数 $N_w$ に応じて10個前後となるように，$N = \lfloor 6 \log(N_w/10 + 1) \rfloor$ として実験的に定めた．

図2は，表1に示す情報科学関連の5つの講義のシラバスから抽出した用語集合をノード，上位下位関係

**Systematization of Course Syllabi Based on Term Sets**
SEKIYA Takayuki† and Kazunori Yamaguchi‡
†Information Technology Center, the University of Tokyo
3-8-1 Komaba, Meguro, Tokyo, Japan
‡Graduate School of Arts and Sciences, the University of Tokyo
sekiya@ecc.u-tokyo.ac.jp

[1] http://mecab.sourceforge.net/

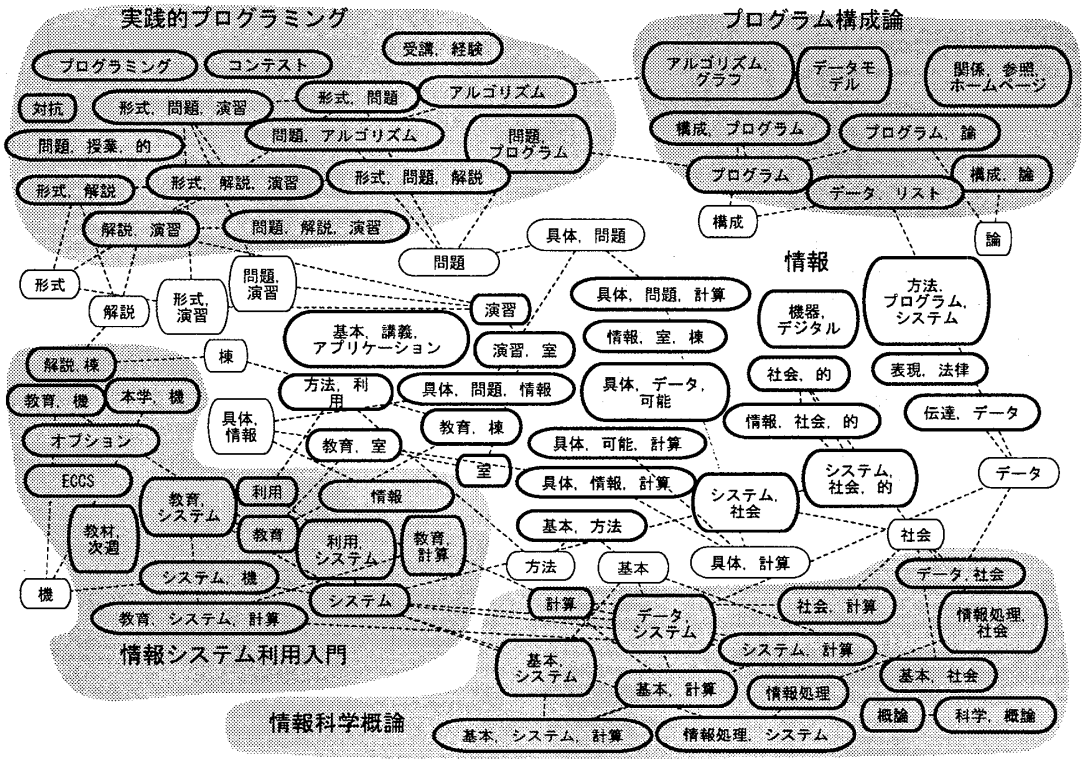


図 2: シラバス構造

表 1: 情報科学に関する講義

| タイトルと内容 (抜粋，一部修正) |
|---|
| **実践的プログラミング**: ACM-ICPC などに出題された問題を題材として、アルゴリズムを考えてプログラムを作る能力を実践的に養うことを目指す. |
| **プログラム構成論**: プログラミングの基礎となるデータモデルについて学ぶ。 |
| **情報**: 多様な場と状況で情報処理システムと関わる全ての人々に必要となる，情報の本質とその人間的および社会的な側面を理解する能力を身に付ける. |
| **情報システム利用入門**: 本学の教育計算機システムを利用して，一般的な情報システムの利用方法を初歩から学ぶ. |
| **情報科学概論**: 現代社会の基盤構造の一つになりつつある情報処理システムおよびそれが扱うデータについての，系統的な枠組の理解を目的とする. |

をエッジとしたグラフである．同じ講義のシラバスから抽出された用語集合を図形で囲み，講義同士の関係構造を表現している．講義「情報」と「情報システム利用入門」が用語集合 { 情報 } を共有したり,「プログラム構成論」と「実践的プログラミング」で { アルゴリズム } の上位下位関係等で，相互に関連のある講義であることが読み取れる．

一方,「情報」と「プログラム構成論」が実際には強く関連があるにも関わらず，図 2 では両講義の用語集合を繋ぐエッジが比較的少なく見える．これは「情報」のシラバスが毎回の授業内容の詳細を記述しているのに対して,「プログラム構成論」のシラバスが全体概要を記述するのみで，両者の記述の粒度が異なっており，抽出された用語集合に違いがあったためである.「プログラム構成論」でも詳細な内容を記述するなど，シラバスを見直すべきであると解釈できる.

## 4 終わりに

シラバスに含まれる用語の集合に基づく，講義同士の関係構造を分析する方法とその分析結果を示した．今後，関係構造の可視化方法を開発する予定である.

## 謝辞

本研究の一部は科研費 (17200048) の助成を受けて実施したものである.

## 参考文献

[1] T. Sekiya and K. Yamaguchi. Knowledge systematization of online syllabus and curriculum design. In *ITHET 2003, 4th International Conference on Information Technology Based Higer Education and Training*, pp. 343–347, 2003.

[2] Gerald Salton. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley, Massachusetts, 1989.