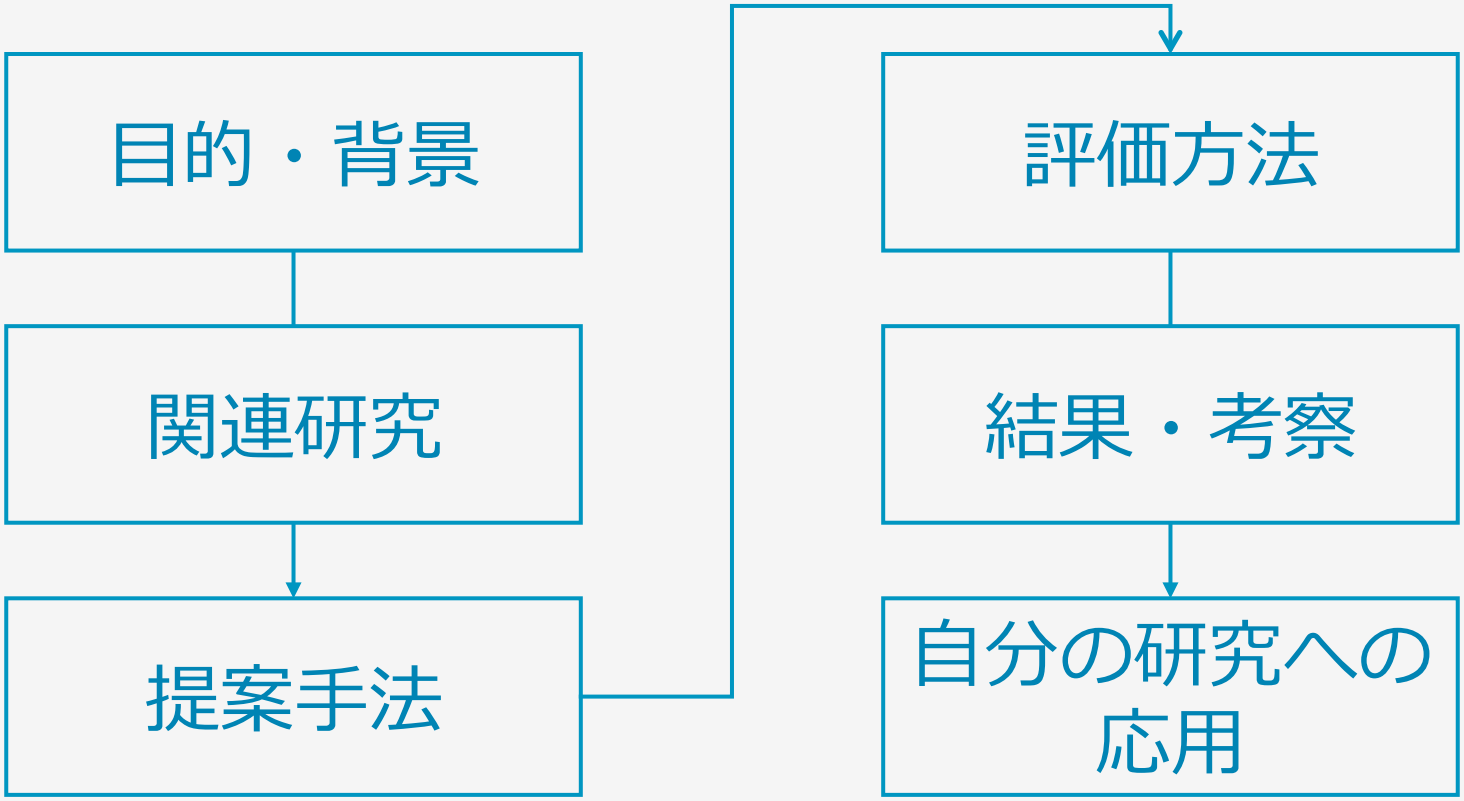
繰り返し構造の検出に基づく Webページの見出しの階層構造の解析

ITシステムプロジェクト

政策・メディア研究科 修士1年
笹本 将平

目次



研究の目的

目的：Webページの見出しの階層構造の解析

1. 検索した時に関係ないページまで出てくるのは問題である。

2. ページ内の見出しの階層構造を用いることで検索エンジンの性能は向上するらしい。

3. Webページ中の繰り返し構造を抽出し、見出しの階層構造をうまく解析したい。

関連研究

[4]：繰返し構造に基づいたWebページの構造化

繰り返し構造を検出し、同じレベルの情報をセグメンテーションすることで、Webページを構造化する。

レイアウトに関する情報は用いていない。

関連研究

[5]：Webページのテキストセグメント階層構造の抽出

教師あり機械学習で親子関係を決定する。
特徴量には

1. DOMのパス
2. インデント情報
3. 言語情報

を利用する。

文字に関する視覚的な情報は用いていない。

関連研究

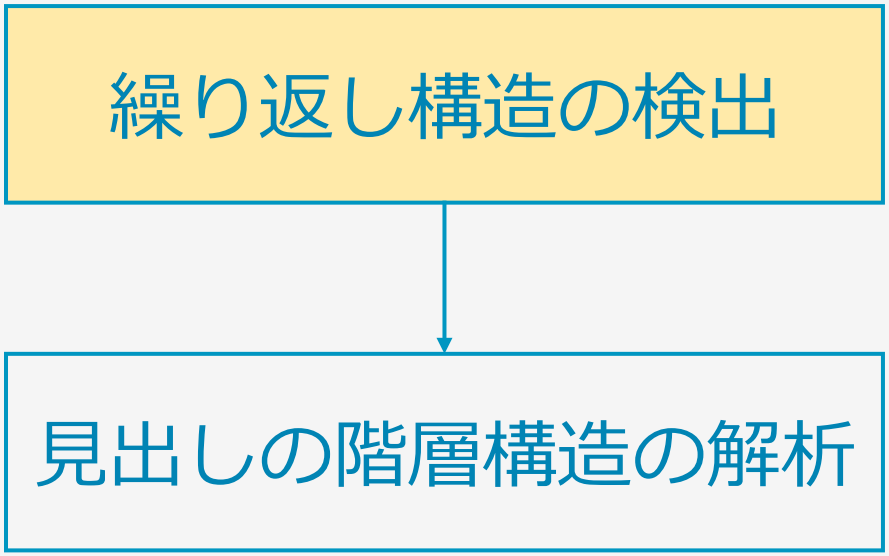
[6]：Detecting Web Page Structure for Adaptive Viewing on Small Factor Divices.

モバイル端末での最適な情報表示のために、
DOM木を手がかりにページを分割する。

DOM構造に依存するため、イレギュラーな構造に対応できない。

提案手法

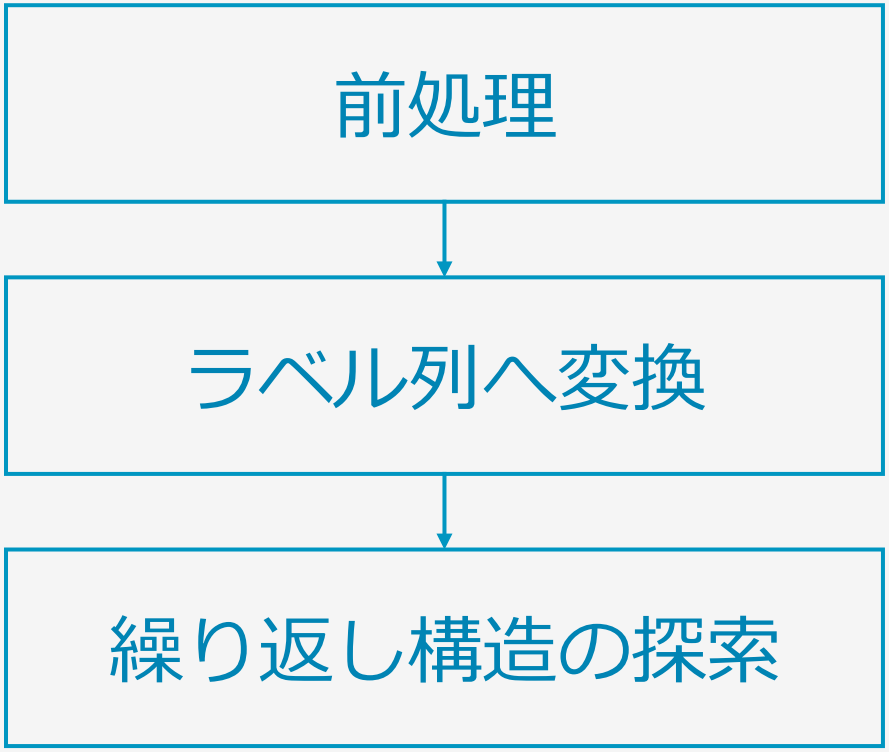
大きく2つのステップに分かれる



本論文では、繰り返し構造の検出手法の改善が主題。

提案手法

繰り返し構造の検出方法



サンプル


慶應義塾
一般の皆様
受験生の皆様
在校生の皆様
卒業生の皆様
ご支援をお考えの皆様
慶應義塾の紹介
教　育
研　究
医　療
社会貢献
国際連携
ピックアップ
学部入学案内
図書館
慶應義塾の情報公開
研究業績・教員情報
学生生活・奨学金
就職・進路
慶應義塾を知る・楽しむ
コンテンツ一覧
半学半教（06/09更新）
New
大学でのゼミのご紹介を通して、教員と学生による教育・研究の取り組みを紹介します。（理工学部数理科学科　栗原将人研究室）
大学資料請求
東日本の復興に向けて
支援活動／対応

前処理

1. HTMLファイルを整形
   - 省略された終了タグの補完
   - 開始タグ、終了タグの対応関係を修正
2. 表を記述するのに用いられたtableタグにラベルTを付与する（自動判別）

サンプル

前処理されたHTMLファイル

ラベル列への変換

1. HTMLファイルをパースし、タグ以外のテキストを行単位で区切り、IDを振る
   - 行単位とはHTMLタグで分割される単位のこと
2. 同時に各行の属性情報を取得する
3. 同じ属性を持つ行同士で1つのグループを為すように全ての行をグループに分類する

ラベル列への変換

4. 全グループに重複しないように、
   ラベル（0,1,2,3…）を付与する
   - セパレータと思われる要素には
     ラベルSn（n=0,1,2,3…）をつける

サンプル

行 I D	属 性	ラ ベ ル
0	foo, bar	0
1	foo, bar, baz	1
2	foo, bar	0
3	baz	T
4	bar	S_0

繰り返し構造の探索

セパレータありの繰り返し構造
を検出

↓

セパレータなしの繰り返し構造
を検出

セパレータありの場合

セパレータを1つの要素と考えると精度が落ちるので、先に処理しておく。

1. セパレータの前後のブロックを比較
2. 条件が一致したら、前後のブロックは同じ繰り返し構造の要素
   - 条件：リンクかどうか、色、背景色の3つ
3. ラベルNx（x=0,1,2,3…）を付与する

セパレータなしの場合

基本処理

① ラベル列 X 内において 2 回以上出現する，“T”，“Nx” 以外のラベルの集合を L={L1,L2,L3,...Ln}，繰り返し構造を構成するブロックの集合を R={},X 内のラベルを指すポインタを i とする．

② ラベル集合 L からあるラベルを取り出す．取り出されたラベルを Lx とする．

③ ポインタ i を X の先頭にセットする．

④ i を一つずつずらしてゆき，Lx が出現した位置をブロック候補の先頭とする．

⑤ i を一つずつ後方へずらしてゆき，「Lx が次に出現する箇所の直前」か「R に含まれるブロックの先頭の直前」か「R に含まれるブロックの末尾」までをブロック候補の末尾とする．ただし，ブロック候補に含まれるラベル数が 1 の場合，4.2 で述べた見出し特徴付き要素でなければ，そのブロック候補は破棄する．

⑥ i を「ブロック候補の末尾＋1」にセットする．

セパレータなしの場合

基本処理

⑦ ④～⑥の処理をiがXの末尾にたどり着くまで繰り返し，ブロック候補群を作成する．

⑧ ②～⑦で作成されたブロック候補群をRに追加する．

⑨ ②～⑧の処理を，Lが空になるまで繰り返す．

⑩ 最後に探索できたブロックから最初に探索できたブロックへ逆順で各々と隣接しているブロックとの類似度を計算し，繰り返し構造を構成しうるかどうかの判定を行う．構成しうると判断された場合は繰り返し構造プールに保存する．ただし，類似度を計算する前に以下の処理を行う．これは，4.3 で述べたパターンの反復回数の差異を吸収するためである．(1)繰り返し構造プールを参照し，比較するブロック中に繰り返し構造があれば1つにまとめる．(2)同一のラベルが連続している箇所が比較するブロック中にあれば1つにまとめる．

セパレータなしの場合

処理1～9

初期状態：(1, 2, 3, 1, 2, 3, 1, 2, 4, 3)

↓

1に着目：(1, 2, 3, 1, 2, 3, 1, 2, 4, 3)

↓

ブロック候補群Rに追加

↓

2に着目：(1, 2, 3, 1, 2, 3, 1, 2, 4, 3)

⋮

繰り返し構造プールへの保存

ID	先頭要素のID	末尾要素のID	基礎単位
0	3	9	1,2,3
1	0	5	1,2,3
2	?	?	???

評価方法

- 評価用テストセット
  - 200ページを対象
  - 被験者3名に見出しとその支配範囲を抽出し、判断が一致した166ページをテストセットとする
  - この166ページに対し、人手で見出し、支配範囲、繰り返し構造に関する情報を付与する
  - クローズドテストセット49ページ
    オープンテストセット117ページに分ける

繰り返し構造の検出精度

表 5　繰り返し構造の抽出実験（先行研究）

	PERFECT	FA	MISS	Precision	Recall	F-Measure
ClosedTest	98	119 （中の 70 ファイル部分一致）	186 （中の 70 ファイル部分一致）	0.452	0.345	0.391
OpenTest	199	296 （中の 161 ファイル部分一致）	462 （中の 161 ファイル部分一致）	0.402	0.301	0.344

表 6　繰り返し構造の抽出実験（提案手法）

	PERFECT	FA	MISS	Precision	Recall	F-Measure
ClosedTest	114	152 （中の 75 ファイル部分一致）	140 （中の 70 ファイル部分一致）	0.429	0.449	0.438
OpenTest	252	354 （中の 163 ファイル部分一致）	334 （中の 137 ファイル部分一致）	0.416	0.430	0.423

先行研究より精度が23%あがった。

見出しの抽出精度

表 7　見出しの抽出実験（先行研究）

	HIT	MISS	FA	Precision	Recall	F-Measure
ClosedTest	1381	537	531	0.722	0.707	0.714
OpenTest	2659	1207	1226	0.684	0.688	0.685

表 8　見出しの抽出実験（提案手法）

	HIT	MISS	FA	Precision	Recall	F-Measure
ClosedTest	1567	265	441	0.780	0.855	0.816
OpenTest	3411	928	1001	0.773	0.786	0.780

先行研究より精度が14%あがった。

見出しの階層構造の検出精度

表 9　見出し二項関係抽出実験（先行研究）

	HIT	MISS	FA	Precision	Recall	F-Measure
ClosedTest	2598	1513	1088	0.705	0.632	0.670
OpenTest	3876	3131	2331	0.624	0.553	0.586

表 10　見出し二項関係抽出実験（提案手法）

	HIT	MISS	FA	Precision	Recall	F-Measure
ClosedTest	2709	1409	1102	0.711	0.658	0.683
OpenTest	5519	3626	3180	0.634	0.603	0.619

先行研究より精度が6%あがった。

考察

- 階層構造をうまく判定できない場合、
  多くは見出しの抽出の段階での誤りが原因であった。
- よって、見出しの抽出能力を向上させる
  ことが精度の向上につながるのではないか。

まとめ

- Webページの繰り返し構造の検出方法を改善、精度を向上させた。
- 見出し階層構造の解析においても、先行研究より良い結果が得られた。

FIN.

ご静聴ありがとうございました

SAMPLE SLIDE

- 発表用に調整されたシンプルデザイン

1. **フォント**

   和文フォントをメイリオ、欧文フォントを
   Segoe UI にデフォルト設定。

2. **カラー**

   黄色と水色の彩度を落とし、明るい印象を
   残したまま、落ち着いた配色設定。

3. **見本付き**

   オブジェクトやテキストの実際の配置や
   デザインの見本用スライドがあります。

比較ボックス

Aについて

- SAMPLE
  - A
  - A

Bについて

- SAMPLE
  - B
  - B

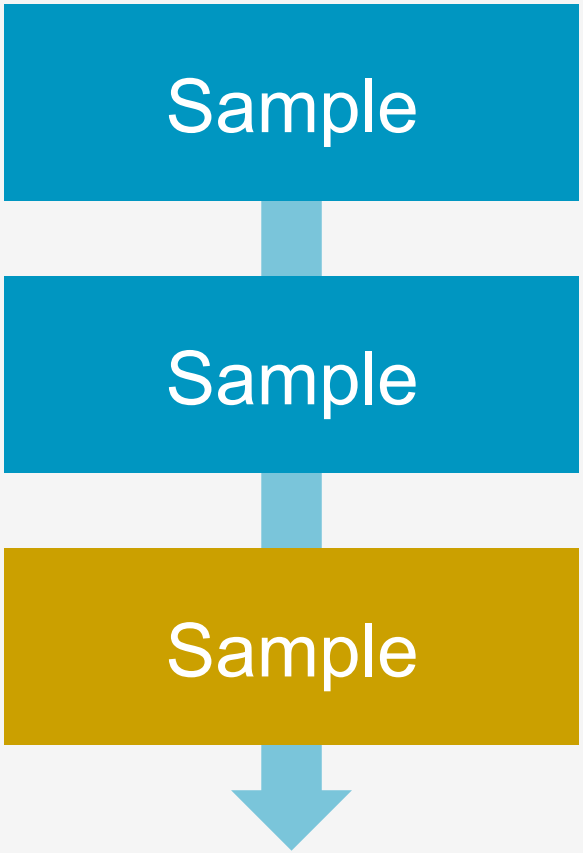
中表紙

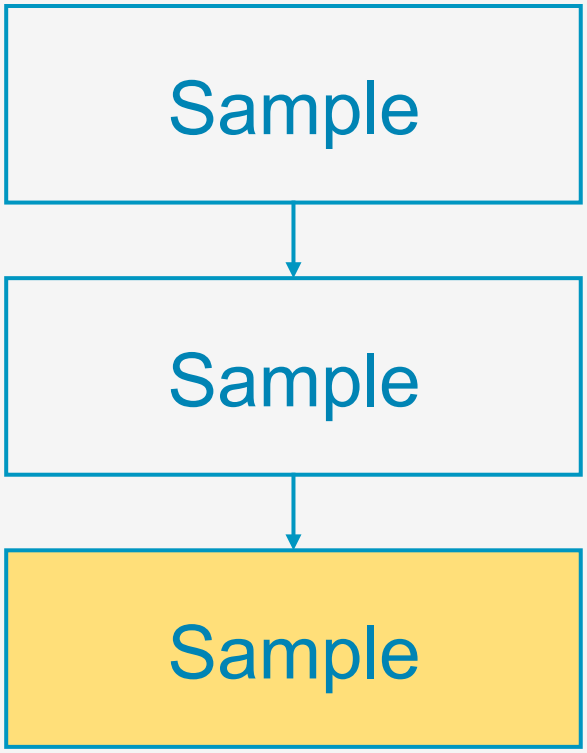
セクションの区切りなどに

見本｜通常の箇条書きスタイル

- 第 1 項目のサンプル
  - 第 2 項目のサンプル
    - 第 3 項目のサンプル
    - 第 3 項目のサンプル
- **第 1 項目のサンプル**
  - 第 2 項目のサンプル
    - 第 3 項目のサンプル
    - 第 3 項目のサンプル

見本｜オブジェクト

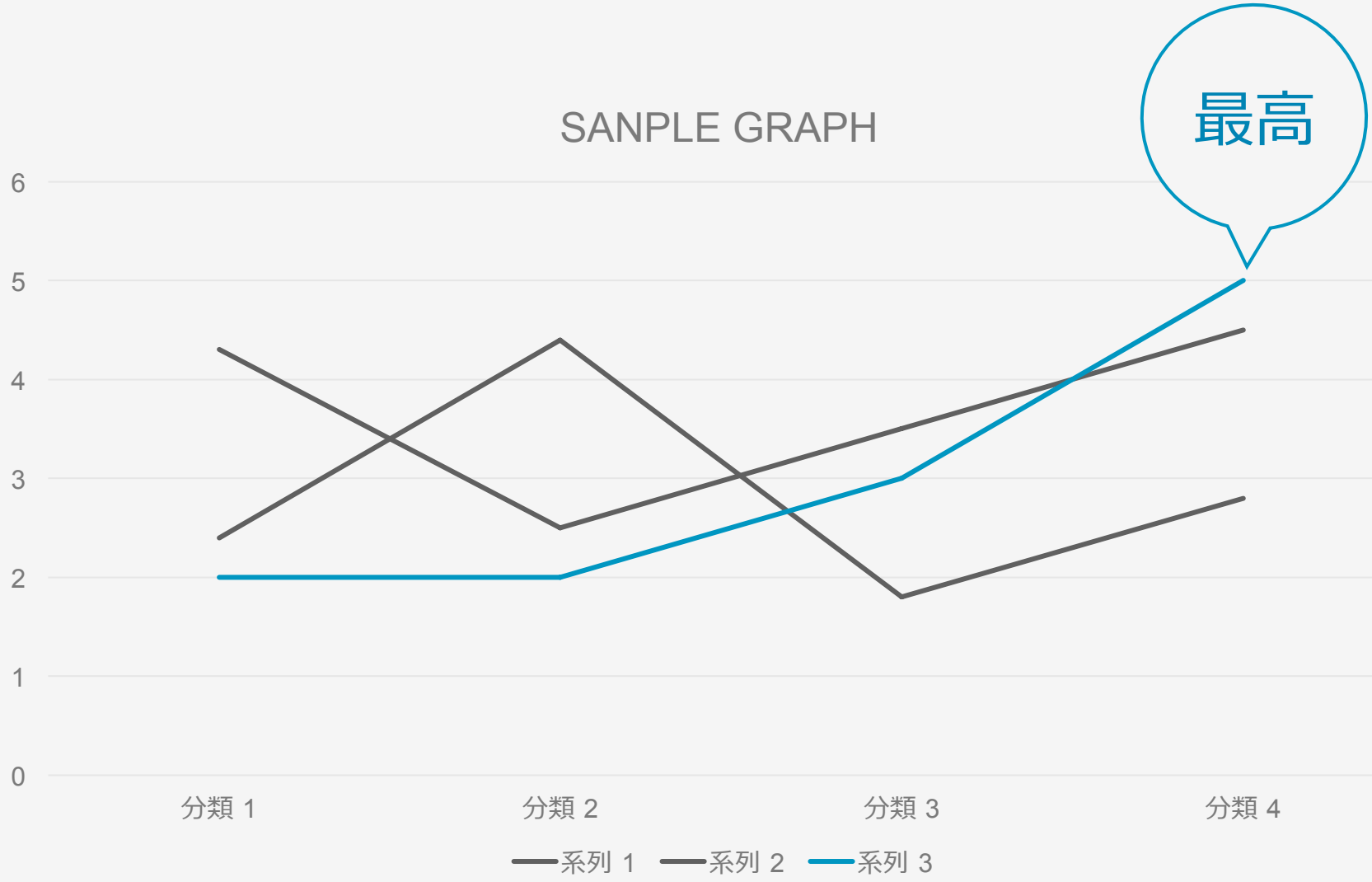




見本 | 表

素 材	器 具	結 果
素材01	器具X	0.01%
素材01	器具Y	0.33%
素材01	器具Z	0.03%
素材02	器具X	0.95%
素材02	器具Y	0.22%

見本 | グラフ



見本｜文章の見せ方の例

見出しはこのような感じで

本文はこのような感じで書き、強調する時は
太字にするか、<u>下線を引く</u>ようにする

- 箇条書きも同様にする
- 特に重要な用語には色もつける
- 文章はなるべく位置をそろえ、
  色は水色と黄土色を使いまわすように

まとめはこのような感じで堂々と