Webページタイプによるクラスタリングを用いた検索支援システム

折原　大　　内海　彰
電気通信大学　システム工学専攻

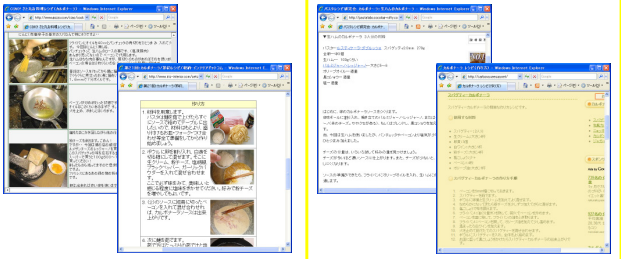
はじめに

- 背景
  - 文書クラスタリングを用いた検索支援システム
    - Clusty(http://clusty.jp/)
    - KartOO(http://www.kartoo.com/)
    - Carrot(http://www.carrot-search.com/)
  - これらはすべてトピックによる分類を行っている
- 動機
  - ユーザが望む分類はトピックだけではない
    - ニュースサイト／blogなどジャンルによる分類
    - 画像や動画の有無による分類
    - 企業・大学などのオフィシャルサイトかどうかによる分類

分類例1

例1：カルボナーラのレシピを写真つきで欲しい！

レシピ（画像つき）　レシピ（文字のみ）



分類例2

例2：年金問題についてのニュース記事／個人的な意見が知りたい！

ニュースサイト　blogサイト

本研究の目的

- 本研究の目的
  - HTMLタグを用いることで，トピックによる分類ではなく，Webページの形式（ページタイプ）による分類
  - 用意されたカテゴリへの分類(classification)ではなく，クラスタリング手法を用い検索結果に応じた動的な分類(clustering)
  - HTMLタグの出現頻度情報を元にした素性の提案

関連研究との比較 - 分類手法

- トピックによる分類
  - 予め用意したカテゴリへの静的な分類(classification)
    - 同義語，多義語の考慮による文書分類の精度向上 [上嶋,04]
  - クラスタリングによる動的な分類(clustering)
    - 構造的言語処理による大規模ウェブ情報のクラスタリング [馬場,07]
    - A Search Result Clustering Method using Informatively Named Entities [Toda,05]
- ページタイプによる分類
  - 予め用意したカテゴリへの静的な分類(classification)
    - Learning to Classify Documents According to Genre [Finn,03]
    - Multiple Sets of Features for Automatic Genre Classification of Web Documents [Lim,05]
  - クラスタリングによる動的な分類(clustering)
    - Unsupervised Non-topical Classification of Documents [Bekkerman,06]（note:新聞記事を対象としている）

⇨本研究ではWebページタイプによるクラスタリング手法を提案

関連研究との比較 - 素性

- 関連研究で扱われている素性
  - 語に基づく情報
    - 単語の出現頻度(Bag-of-Wards, BoW)
    - 品詞の出現頻度(Part-of-Speech, PoS)
    - 各カテゴリに固有のキーワード
  - 文書に基づく情報
    - 疑問文，命令文などの文型や，名詞句や動名詞句などの句の出現頻度
    - 文や段落の平均の長さなどの統計的情報(Text Statistics)
  - Web特有の情報
    - HTMLタグの出現頻度
    - タイトルに関する情報
    - URLに関する情報(深さ，ドキュメントタイプ(html,pdfなど)，ドメインなど)

⇒本研究ではHTMLタグの出現頻度を元にした関連研究とは異なる新しい素性を提案

6

ページタイプによるクラスタリングを用いた検索支援システム

1. Live Searchより検索結果上位n件を取得
2. 各ページのHTMLソースを取得
3. 次の3つのStepでクラスタリングを行う
   - Step-1 特徴ベクトルの構成
     - Step-1F HTMLタグの頻度に基づく特徴ベクトル
     - Setp-1T HTMLタグの木構造に基づく特徴ベクトル
   - Step-2 類似度の計算
   - Step-3 クラスタの生成
4. 各クラスタの重心に最も近いページをクラスタの代表とし，キャプチャ画像をユーザに提示

7

Step-1F 頻度に基づく特徴ベクトル

- 各WebページをHTMLタグの頻度に基づく特徴ベクトルで表現
  1. HTMLタグを抽出
  2. 「分割数」と「n-gram」による特徴ベクトルの属性を決定
  3. 「属性値のカウント方法」と「IDF値の考慮の有無」による属性値を計算
  4. 各特徴ベクトルの長さを1に正規化

8

Step-1F.2 属性の決定

- 分割数
  - タグがどの位置に出現しているかを考慮する要素
  - 抽出されたタグを分割数mで等分し、各範囲で1つの属性とみなす
- n-gram
  - 連続するタグの組み合わせを考慮する要素
  - 抽出されたタグを連続するn個の組み合わせで1つの属性とみなす

9

Step-1F.2 属性の決定(分割数)

HTML Document

<H1>
<H2>
<P>
<P>
....
<H2>
<P>
<OL>
<LI>　2つに分割
<LI>
<LI>
....
<H2>
<P>
<P>

A
B

- 分割数が2の場合

	A	B
<H1>	1	0
<H2>	2	1
 	1	0
<P>	3	2
<OL>	1	0
<LI>	1	2

10

Step-1F.2 属性の決定(n-gram)

HTML Document

<H1>
<H2>
<P>
<P>
....
<H2>
<P>
<OL>
<LI>
<LI>
<LI>
....
<H2>
<P>
<P>

<H2><P>
<P><P>

- n-gramが2の場合

<H1> 	1
 <H2>	1
<H2><P>	3
<P><P>	2
…	

11

Step-1F.2 属性の決定

HTML Document → 2-gram

<H1>
<H2>
<P>
<P>
....
<H2>
<P>
<OL>
<LI>
<LI>
<LI>
....
<H2>
<P>
<P>

<H1>

<H2>
<H2><P>
<P><P>
....
<H2><P>
<P><OL>
<OL><LI>
<LI><LI>
<LI>....
....
<H2><P>
<P><P>

A
B

- 分割数が2,かつ, n-gramが2の場合

	A	B
<H1> 	1	0
 <H2>	1	0
<H2><P>	2	1
<P><P>	1	1
...	...	...

Step-1F.3 属性値の計算

- 属性値のカウント方法
  - 一般的な出現回数をカウントする「頻度」
  - その属性が出現したかどうかの2値をとる「有無」
- IDF値の考慮の有無
  - IDF値の考慮「あり」
  - IDF値の考慮「なし」

Step-1F.3 属性値の計算(頻度・有無)

HTML Document

<H1>
<H2>
<P>
<P>
....
<H2>
<P>
<OL>
<LI>
<LI>
<LI>
....
<H2>
<P>
<P>

- 「頻度」と「有無」

	「頻度」	「有無」
<H1>	1	1
<H2>	3	1
<H3>	0	0
<P>	5	1
<OL>	1	1

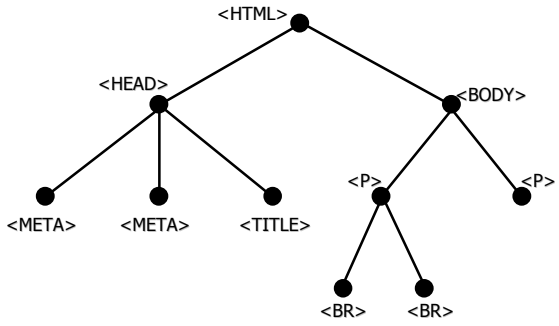
Step-1T 木構造に基づく特徴ベクトル

- 各WebページをHTMLタグの木構造に基づく特徴ベクトルで表現
  1. HTMLタグの木構造を2分木に置き換える
  2. 2分木に対応するBinary Branchを定義する
  3. Binary Branchを用いてBinary Branch Vectorを求めこれを特徴ベクトルとする
  4. 各特徴ベクトルの長さを1に正規化する

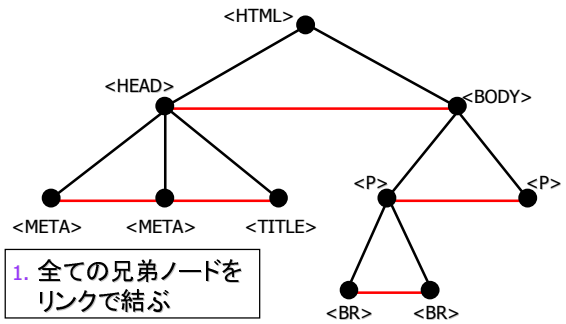
Step-1T.1 2分木へ置き換え

- HTML文書からHTMLタグの木構造を取り出し、次の方法で2分木へ置き換える
  1. すべての兄弟のノードをリンクで結ぶ
  2. 各ノードの最初の子ノードとのリンクを除く全てのリンクを削除する
- 変換後に該当する子ノードがない場合はノードεを付加する

Step-1T.1 2分木へ置き換え

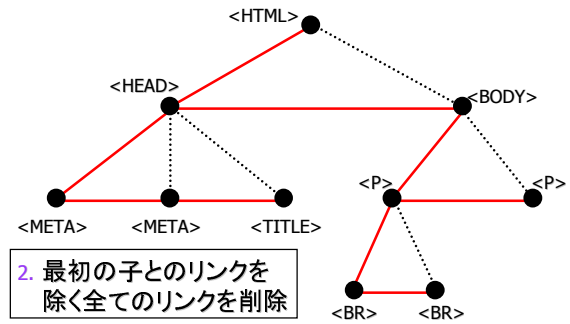


Step-1T.1 2分木へ置き換え



1. 全ての兄弟ノードを
リンクで結ぶ

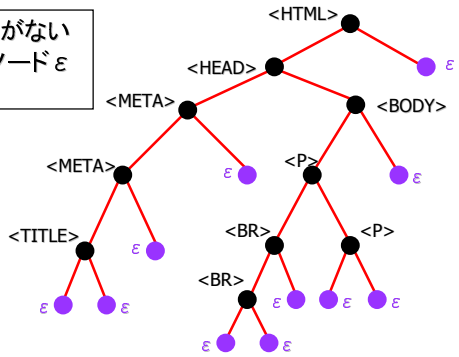
Step-1T.1 2分木へ置き換え



2. 最初の子とのリンクを
除く全てのリンクを削除

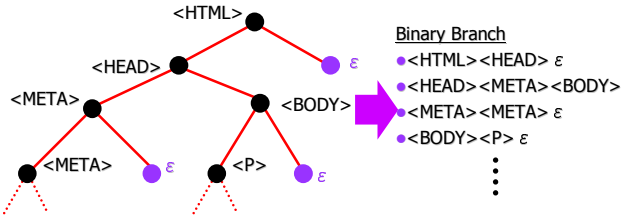
Step-1T.1 2分木へ置き換え

☞ 子ノードがない場合はノードεを付加



Step-1T.2 Binary Branchを定義

- Step-1.1で作成された2分木のうち、1階層分を取り出したものをBinary Branchとする



Step-1T.3 Binary Branch Vector

- Step-1.2で求めたBinary Branchを要素とし、各要素の値は頻度とするBinary Branch Vectorを求める
  →これを特徴ベクトルをする

Binary Branch					
	<HTML> <HEAD> ε	<HEAD> <META> <BODY>	<META> <META> ε	<BODY> <P> ε	…

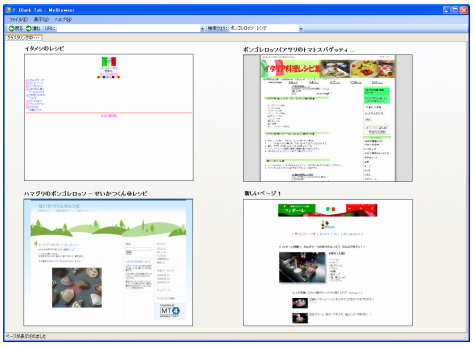
特徴ベクトル ＝（ 1, 1, 1, 1, … ）

Step-2 類似度の計算
Step-3 クラスタの生成

- 類似度
  - 多次元ユークリッド空間の距離を利用
- クラスタリング手法
  - クラスタリングアルゴリズム：階層的手法
  - クラスタ間の類似度の計算手法：Ward法
  - 停止条件：ページ総数の4割を超えるクラスタが作成される直前

検索支援システム　出力例

- C#により作成



評価実験

- 提案する手法を実装し，有用性を検証
  - 分類精度による評価
    - データ
      - アンケートにより作成した分類正解データ(21件)
    - 比較手法
      - 単語の分布に基づく手法(BoW)
      - Bekkermanらの手法[Bekkerman,06]
  - 検索支援システムとしての評価
    - データ
      - 2名のユーザに試用してもらい，回答となるページを取得するまでの早さ，多さを比較
    - 比較手法
      - Live Search による検索と比較

評価データ - 分類精度 (1/3)

- 以下の手順で正解データを作成
  1. 各人が検索エンジンを用いて自由に検索
  2. 得られた検索結果の上位100件を全て閲覧
     - PDF, XMLなどは対象外とする
     - 分類が難しいページは「その他」に分類してもらい、評価データからは対象外とする
  3. 「見た目やスタイルが似ているものどうしに分類してください」と教示し、2.で閲覧したページを自由に分類

評価データ - 分類精度 (2/3)

表1：アンケートにより作成した評価データのページ数およびグループ数

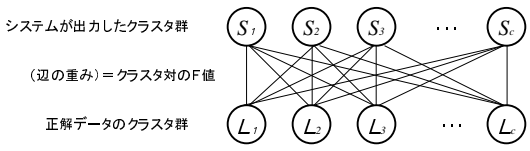
	ページ数	グループ数		ページ数	グループ数
Data01	55	4	Data12	43	5
Data02	47	8	Data13	51	8
Data03	49	6	Data14	40	3
Data04	44	5	Data15	74	3
Data05	51	9	Data16	93	9
Data06	36	3	Data17	47	4
Data07	46	4	Data18	99	6
Data08	35	5	Data19	68	3
Data09	45	7	Data20	54	3
Data10	44	4	Data21	56	3
Data11	43	6			

評価データ - 分類精度 (3/3)

- 正解データ例
  - Date07
    - 検索クエリ：「最近，人気，映画」
    - ユーザが付けた分類グループ名
      - 映画関連のニュースサイト
      - 映画の内容，人物などの紹介
      - 映画製品DVDなどの紹介
      - ブログなどの個人の意見，感想
  - Data21
    - 検索クエリ：「ロボット，学習，制御」
    - ユーザが付けた分類グループ名
      - 学校機関系
      - 書籍関係
      - 解説系

評価基準 - 分類精度

- F値の考え方をもとに、クラスタ群対のF値を計算
  - 完全2部グラフの重みつき最大マッチング問題を解くことでクラスタ群対のF値とする



  - ここではシステムが出力するクラスタ数は正解データのグループ数と同数とする

評価結果 - 分類精度 (1/2)

- HTMLタグの頻度に基づく特徴ベクトルの構築では，以下のパラメータが最適
  - 分割数：3
  - n-gram：2
  - 属性値のカウント方法：有無
  - IDF値の考慮の有無：なし

表2：属性値，IDFの考慮の違いによる平均F値

		属性値	
		頻度	**有無**
IDFの考慮	あり	0.444	0.445
	なし	0.444	**0.450**

表3：n-gramによる平均F値

n-gram	平均F値
1-gram	0.460
2-gram	**0.462**
3-gram	0.457
4-gram	0.438
5-gram	0.433

表4：分割数による平均F値

分割数	平均F値
1	0.457
2	0.460
3	**0.477**
4	0.454
5	0.464

評価結果 - 分類精度 (2/2)

- 比較手法よりも本研究で提案する2つの手法において分類精度が向上

表5：提案手法と既存手法との比較

	平均F値
タグの木構造に基づく特徴ベクトル	**0.478**
タグの頻度に基づく特徴ベクトル（最適なパラメータ）	**0.477**
Bekkermanらの手法	0.459
Bag-of-Words (BoW)	0.451

評価結果 – 検索支援システム

- 2名のユーザに試用してもらった
  - 次のような検索要求において本システムが有用であった
    - 料理のレシピを検索した際に，画像付きで解説されているページが欲しい
    - 文書クラスタリング手法を検索した際に，具体的な内容が書かれているページが欲しい

      ⇒学会のプログラムが書かれているページが分別された
- 今後，検索要求タスクを設定し本評価を行う

今後の課題

- 検索支援システムとしての問題点を改良
  - 検索結果（クラスタリング結果）出力までの時間がかかりすぎる
    - 30件の検索結果をクラスタリングするのに約1′30″
  - クラスタリング結果の提示方法
    - クラスタの代表となるページのキャプチャ画像を提示しているが…
- トピックとページタイプを組み合わせたクラスタリング手法の提案