修 士 論 文

アンカーテキストを用いた
属性情報の抽出

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

太田 茂

2007年3月

修 士 論 文

アンカーテキストを用いた
属性情報の抽出

指導教官　鳥澤健太郎 助教授

審査委員主査　鳥澤健太郎 助教授
審査委員　東条敏 教授
審査委員　白井清昭 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

510019 太田 茂

提出年月: 2007 年 2 月

概 要

本稿では,Web 上の HTML 文書中から標準サイトマップの生成に必要なアンカーテキストの異表記同義語関係をアンカーテキストが指す文書のクラスタリングによって獲得する手法を提案する.

目 次

第 1 章　はじめに **1**
1.1　研究の背景と目的 1
1.2　本論文の構成 2

第 2 章　関連研究 **3**
2.1　属性抽出 3
2.1.1　上位下位関係を用いた HTML 文書からの属性及び属性値の自動抽出 3
2.1.2　Web からの属性情報記述ページの発見 4
2.2　クラスタリング 5
2.3　その他 5

第 3 章　提案手法 **6**
3.1　手法概要 6
3.1.1　本研究におけるサイトマップ 6
3.1.2　標準サイトマップの利用例 9
3.1.3　標準サイトマップ生成時の問題点 9
3.1.4　アンカーテキストの異表記同義語 10
3.2　異表記同義語関係の抽出の流れ 11
3.2.1　Step1:リンク情報の抽出と適切なアンカーテキストの抽出 11
3.2.2　Step2:索引語ベクトルの生成と文書の分類 13
3.2.3　Step3:異表記同義語関係にあるアンカーテキストの同定 14
3.3　システム概要 15
3.3.1　文書収集処理 16
3.3.2　タグ情報抽出処理 16
3.3.3　索引語抽出処理 17
3.3.4　索引語ベクトル生成処理 19
3.3.5　文書のクラスタリング 21
3.3.6　異表記同義語関係の同定 24

第 4 章　実験 **26**
4.1　目的 26

4.2 方法 26
4.2.1 実験手順 26
4.2.2 評価方法 27
4.3 準備 27
4.4 結果 27
4.4.1 予備実験 27
4.4.2 異表記同義語抽出 30
4.4.3 各種パラメータの強調効果 35
4.5 考察 49

第 5 章 おわりに **50**
5.1 まとめ 50
5.2 今後の課題 50

謝辞 **52**

図 目 次

3.1 サイトマップの表現例 7
3.2 sitemap1 8
3.3 sitemap2 8
3.4 異表記同義語関係 10
3.5 システム概要 15
3.6 フレーム対応文書からの抽出例 17
3.7 全文書数算出のイメージ 21
3.8 初期セントロイドをランダムに決定 23
3.9 初期セントロイドを遠い順に決定 24

4.1 クラスタリング毎の平均 F 値の散らばり (5byouin)) 28
4.2 クラスタリング毎の F 値の散らばり (10byouin) 29
4.3 最適クラスタリング結果 (5lab, research) 32
4.4 クラスタリング結果 1 34

第1章　はじめに

1.1　研究の背景と目的

本論文では, 与えられたクラス (例. 大学等) の対象物 (例. 北陸先端大, 金沢大) を記述した Web サイト (Web 文書の集合) から, 対象物の属性 (例. アクセス, 入学情報) に相当するアンカーテキストの異表記同義語関係を獲得する手法を提案する.

近年, Web の発展に伴い, 病院や大学など, 様々なクラスの対象物について Web サイト (公式ホームページ) が作成され, ユーザが Web から対象物に関する情報を得る際に重要な情報源となっている.

しかしながら, 同じクラスのいくつかの対象物に関する情報を横断的に確認・比較したいと思っても, 各 Web サイトで目的となる情報を含む文書が異なるアンカーテキストで参照されているため, 目的の情報を発見することは難しくなっている. 例えば, 大学のクラスに含まれるいくつかの対象物のサイトから「所在地」に関する情報を知りたいとする. そこで, ユーザはいくつかの対象物のサイトを訪問し「所在地」に関するリンクを辿ることになるが, 「交通案内」や「アクセス」, 「地図」など, 対象物の作成者により異なるアンカーテキストでリンクされている場合も多く, 知りたい側面 (以下, 属性と呼ぶ) について記述されたページかどうかの確認に無用の労力が費やされる.

そこで本研究では, 同一クラスに属する対象物であれば, それらを記述した Web サイトは同じような対象物の属性 (例. 所在地, 交通案内) に関するページ (以下, 文書) を含むという仮定のもと, Web サイト中でそれらの文書を参照するアンカーテキストの異表記同義語関係 (例. アクセス $leftrightarrow$ 交通案内) を獲得することを目指す. ここで, Web サイト中で文書が対象物の属性毎にまとめられているならば, その文書を指すアンカーテキストは, その対象物の属性を端的に表した具体的な単語 (属性語) となっており, これは対象物の属性語の言い換えを獲得することに相当する. この異表記同義語関係を獲得することができれば, 対象物毎に異なるリンク間のラベルを共通の属性語で表示させたそのクラスの標準サイトマップを作成することも可能であり, その標準サイトマップと各対象物の Web サイトとの対応を取ることで, ユーザは標準サイトマップを通して知りたい対象物の属性の情報に容易にアクセスすることが可能となる. また, サイト作成者としても, 個々の対象物のクラスの標準サイトマップを利用することで, そのサイトの視認性を上げユーザビリティ, アクセシビリティの向上を計ることも可能である.

本研究では, 与えられたクラスとその対象物に関する Web サイト (Web 文書の集合) に対し, 以下の手順でアンカーテキストの異表記同義語関係を獲得する.

まず, 各 Web サイト (Web 文書の集合) からリンク情報 (アンカーテキストと参照先の文書のペア) を抽出し, 属性語として適切なアンカーテキストを抽出する (Step1). Step1 で抽出されたペアに含まれる各文書について索引語ベクトルを生成し (Step2), 非階層型クラスタリング手法の k-means 法により文書の分類 (Step3) を行う. その分類結果をもとに, 異表記同義語関係にあるアンカーテキストを同定する. 以下, 各 Step について述べる.

Step1 Web サイトのトップページに相当する文書を解析しアンカータグで示されるアンカーテキストと参照先 URL のペア (リンク情報) を抽出する. 別途, タイトル, メタ情報, 本文を抽出する. トップページから抽出した参照先 URL を再帰的に辿り, サイト中の全アンカーテキストとその参照先 URL のペアを抽出する. さらに, 抽出したアンカーテキストに対し属性にはなりにくい文書の除去処理を行う. 特に, 文字列長の長いアンカーテキスト, 異なるドメインを参照しているアンカーテキスト, 「戻る」「こちら」「ジャンプ」などの属性語とはなりにくいアンカーテキストを除去する.

Step2 各文書の本文をもとに, 索引語ベクトルを生成する. その際, ペアのアンカーテキスト, タイトル, メタ情報, 強調タグ, 索引語の出現頻度 ($TF \cdot IDF$ 値と TF 値) で任意に倍率を設定し重み付ける. 索引語は本文を形態素解析器, キーワード自動抽出システムを利用して獲得する. なお, 属性「交通案内」が指す文書などでは, 本文からの索引語が少数もしくは無い文書があり得る. そのため, 文書に相当する URL とアンカーテキストから類義語, 英和辞書などの言語資源を利用し索引語に含める.

Step3 生成した各文書の索引語ベクトルをもとにクラスタリングを行い分類結果から異表記同義語関係の抽出する. クラスタリングは非階層型クラスタリング手法である k-means 法を用いて行う. その際, クラスタ数 (k) は最も少ない対象物の文書数を用いる.

1.2 本論文の構成

本論文の構成は以下の通りである. 第 2 章で属性抽出, クラスタリング関連に分け関連論文を述べる. 第 3 章で Web サイトの文書から異表記同義語関係を抽出する手法を述べる. 第 4 章では実験結果について述べる. 第 5 章ではまとめと今後の課題について述べる.

第2章　関連研究

本章では概念具体物の属性語の抽出, クラスタリング, その他に分けて関連研究を取り上げる.

2.1　属性抽出

徳永ら [2] は, 統計量, 構文パタンによる頻度,HTML タグによる頻度などと, 上位語を用いた属性情報の抽出を行っている. 本研究では HTML タグによる頻度に注目しベクトル空間の要素の重み付けに利用している. 吉永ら [17] は, 表やリスト形式など視覚的に認知し易い形で記述したページに限定し属性情報の抽出を行っている. 本研究では, ベクトル空間の要素に重み付けする際に利用する.

2.1.1　上位下位関係を用いた HTML 文書からの属性及び属性値の自動抽出

概要

徳永ら [2] は,Web 上の HTML 文書中からの単語の上位下位関係を利用して, ユーザが入力した検索語 (対象語) に関する重要な情報である属性, 属性値を自動抽出する手法を提案している. 以下の仮説の元で 4 つのスコアを手がかりとして属性及び属性値の抽出を行っている.

1. 属性は対象語の上位語を含む文書に現れやすく, それ以外の文書には現れにくい.
2. 属性は HTML 文書中で強調されたり, リストや表の要素になり易い.
3. 属性は対象語の上位語との間に, 助詞' の' を介した固有のパタン・係り受け関係を持つ.

また, スコアは以下に注目している.

1. df や idf などの統計量
2. 上位語を含む特定の単語の構文パタンに適合する頻度
3. 上位語の係り受けの頻度
4. HTML タグに囲まれる頻度

手法としては, 対象語の上位語と HTML タグの情報を利用して属性候補集合の獲得を行う (対象語を下位語に持つ上位語を Shinzato ら [28] により提案された獲得手法により Web から獲得している). 次に獲得した要素数の各属性候補の順位付けを行い, 上位を属性として獲得する. さらに精度の高い属性の抽出を行うため, 属性の各スコア素性として Support Vector Machines(SVM) による学習を行い, 構築された属性モデルによって, 属性としてふさわしくない文字列の削除を試みている. 統計量, 構文パタンによる頻度,HTML タグによる頻度, 新聞記事より獲得した係り受け関係の 4 種類の情報と上位語を組み合わせることで抽出を行った結果,79%の精度で属性を抽出することが可能であることを確認している.

本研究との関連

本研究では,Web 文書の本文やタイトルなどの情報から索引語の抽出を行い, ベクトルを生成する段階で HTML タグによる重み付けを行っている. [2] で示している HTML タグにより強調されている箇所を参照することの有効性を考慮し, ベクトル生成で取り入れる.

2.1.2 Web からの属性情報記述ページの発見

概要

吉永ら [17] は与えられた対象物とそのクラスから, 対象物を記述した代表的な (最も多くの情報を含む)Web ページ (属性情報記述ページ) を発見する手法を提案している. 当システムは, 対象物の属性情報記述ページを, そのクラスの属性知識ベースに基づき発見する. そこで, まず属性情報記述ページにおける属性の現れ方を考慮し一般的な教師なし学習によりクラス属性の知識ベースを構築する. 学習の際には, 属性候補から属性として不適切な単語を除くためにサイト頻度を用いている. ユーザからの入力を通常の検索エンジンを用いて対象物を記述した Web ページを絞り込み, クラスの属性知識を用いスコア付けし, 最良のページを発見している. HTML タグ付きの 0.7TB の日本語 Web 文書を収集しクラスの属性知識ベースを構築し評価したところ,69%の精度で被験者の知りたい属性が含まれていたと結論づけている.

本研究との関連

[17] では知識ベースを作成する際に HTML タグと文字修飾に基づくパターンにより生成しており, 属性-属性値関係の抽出に効果を持たせている. 本研究ではその点に着目し, ベクトル生成時の重み付けを行っている.

[17] における属性獲得に用いた HTML タグを表 2.1 に示す.

表 2.1: 属性獲得に用いたHTMLタグ

HTMLタグ:	TD	TH	LI	DT	DD	B	STRONG	FONT	SMALL	EM	TT

(本研究では強調タグとして使用する上でいくつかのタグを加えている.(表3.3参照))

2.2 クラスタリング

Macqueenら[23]によるk-means法は,非階層型のクラスタリング手法の一つであり,与えられたk個のクラスタに分類する学習量子化の最も基本的なクラスタリング手法である. 単純なアルゴリズム(アルゴリズムについては3.3.5で詳細を述べる)で計算させることができるため本研究において文書分類で使用する.

2.3 その他

湯本ら[4][21]は専門用語を専門分野のコーパスから自動抽出する方法を提案している. ある単名詞が複合名詞を形成するために連接する名詞の頻度を用いている.

Avrim[18]らは,独立した二つの学習器を互いの解析結果を正解データと見なし再学習のプロセスを繰り返す手法を提案している(Co-traning). 彼らはweb文書を分類するためのfeatureとしてテキスト中の単語に加えアンカーテキストを用いて互いの最適解に近づけることをねらっている. 本研究においても文書の類似性を計る上で文書中の単語とアンカーテキストに注目する.

第3章　提案手法

3.1　手法概要

本章では本研究の手法について述べる. 手法としては, サイトマップ作成に必要なアンカーテキスト (属性語) の異表記同義語関係をアンカーテキストが指す文書のクラスタリングによって獲得する.

本研究で開発するサイト情報抽出システムは, あるクラスに属するサイトの HTML 文書を入力元とし, 個々の文書に対し次の操作を施す.

Step1 リンク情報の抽出と適切なアンカーテキストの抽出

Step2 索引語ベクトルの生成と文書の分類

Step3 異表記同義語関係にあるアンカーテキストの同定

階層数, 各パラメータの重みについて実験を行い獲得された結果をもとに他のクラスに対し上記の操作を行い, 有用な属性情報が抽出されるか, 実験と評価を行う. 評価方法については後述する.

図 3.5 に生成システムの全体像を示す. はじめにインターネット上から本システムへ文書群を取り寄せる. 入力値は URL(トップページ) とトップページからのリンク数である. 取り寄せられた文書は HTML ソースのタグ解析機能によって, アンカーテキストによって示されるリンク情報, 文書の特徴を指すきっかけとなる強調タグ, 本文の 3 つを出力する. 次のフェーズで 3 つの入力情報から文書毎に索引語ベクトルを生成する. 生成されたベクトルはクラスタリングにより分類され, アンカーテキストの異表記同義語関係の抽出へとつなげる.

3.1.1　本研究におけるサイトマップ

本研究におけるサイトマップは, サイト構造を有向グラフの木構造で表現したものを指す. 対象ドメインのトップページからの参照情報であるアンカータグに含まれているリンク (参照) 情報によりサイト構造を分析することでサイトマップを生成することができる. また, 本研究では同一クラスのサイトから一般的な項目を抽出したものを標準サイトマップと定義する. 例えば, 学校関連のクラスであれば,「大学案内」,「入学案内」,「学生生

活」などの項目は使われる頻度が高く一般性が高いと判断し, 標準サイトマップの属性項目として取り上げる. 逆にニュースの内容など一時的に発生する項目に関しては一般性の乏しいものとして標準サイトマップには取り入れないものとする.

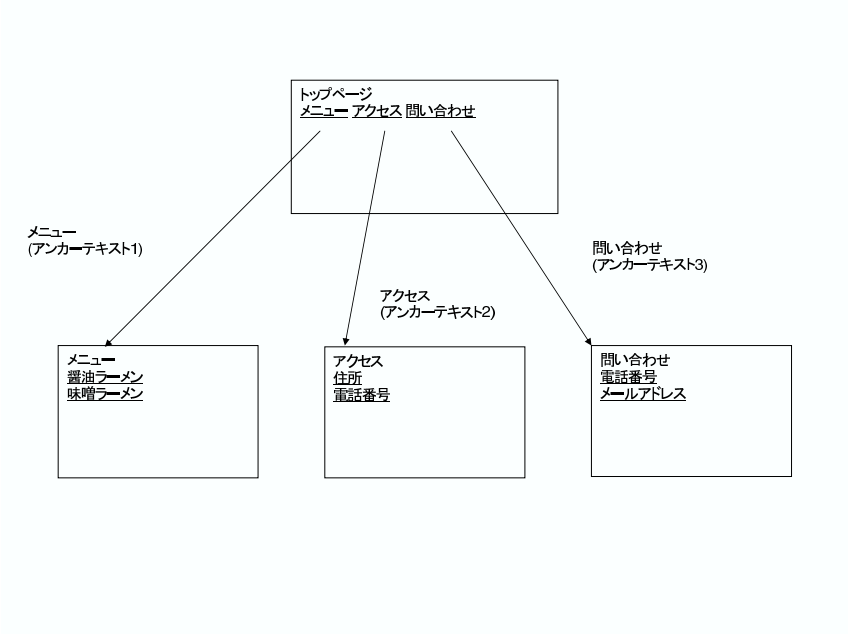

図 3.1: サイトマップの表現例

図 3.1 にサイトマップの表現例を示す. 下線文字がリンクとなっている文字列を指しており, トップページの文書中に「メニュー (アンカーテキスト 1)」「アクセス (アンカーテキスト 2)」「問い合わせ (アンカーテキスト 3)」に関する文書へのリンクが存在し, リンク先のそれぞれの文書からさらに,「醤油ラーメン」「味噌ラーメン」などに関する文書へリンクされている. このように web 文書間でリンクされている関係を人間にわかりやすく表現した一種の地図のようなものを本研究ではサイトマップと呼ぶこととする.

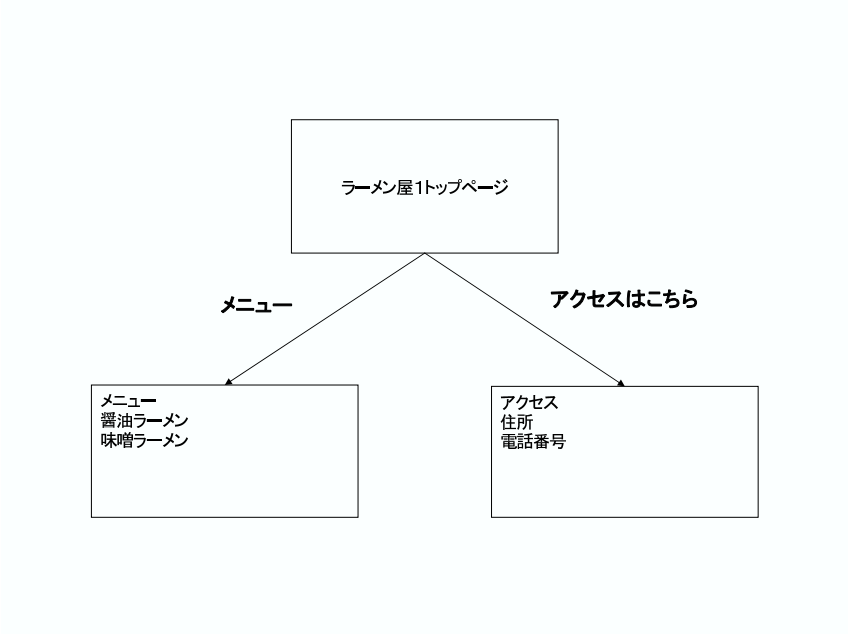

図 3.2: sitemap1

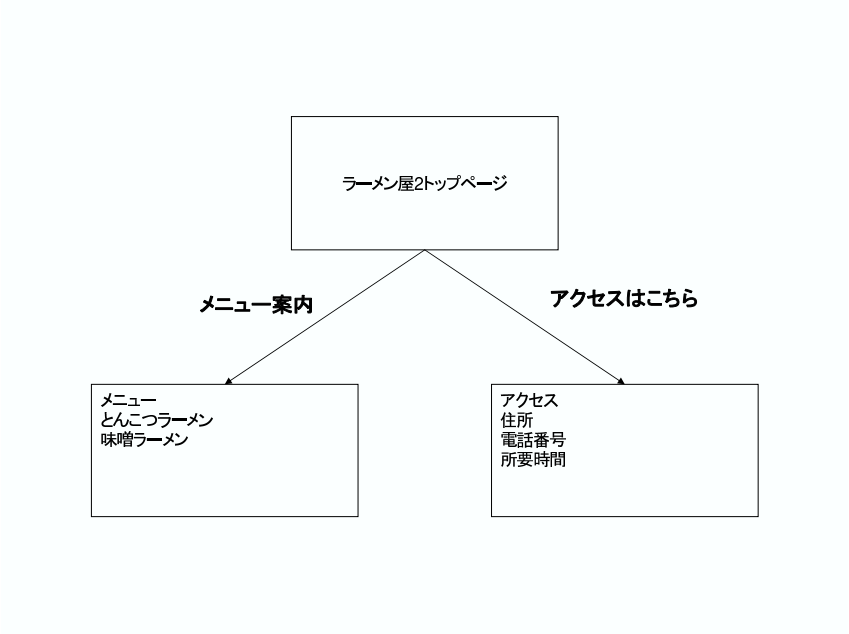

図 3.3: sitemap2

簡易なサイトマップの例を使用し, 同一クラスの簡易なサイトマップを図 3.2, 図 3.3 で示す.

図 3.2 は, ラーメン屋 1 のトップページから「メニュー」というアンカーテキストでメニューに関するページに,「アクセスはこちら」というアンカーテキストでアクセス方法に関するページにリンクが張られている. 一方図 3.3 のラーメン屋 2 では,「メニュー案内」というアンカーテキストでメニューに関するページに,「アクセスはこちら」というアンカーテキストでアクセス方法に関するページにリンクが張られている様子を示している.

本研究は「ラーメン屋 1」の「メニュー」ページと,「ラーメン屋 2」のメニュー案内ページに出現する単語を索引語 (図中の「ラーメン」など) とし異なるサイト同士で類似度の高いページをクラスタリングし, あるクラスタ内のアンカーテキスト同士を異表記同義語関係であることを導く.

3.1.2 標準サイトマップの利用例

本システムにより生成される標準サイトマップの利用例を紹介する. まず一つに既存のサイトを同一クラスの標準サイトマップと統合することで, サイト構成の視認性を上げ, 目的の情報への操作を減らすことができ, アクセス先サイトのユーザビリティ, アクセシビリティの向上を図ることができる. 二つ目に新たなサイトを作成する際の作成方針を決めるきっかけとなるモデルを作成し活用することができる. 同一クラス (ジャンル) の他のサイトの記述項目の集積である標準サイトマップの内容を把握することで標準的な項目や構造を参照することができると考えられる.

3.1.3 標準サイトマップ生成時の問題点

予測されうる問題は以下の 3 つが挙げられる.

属性とみなせないアンカーテキストが多い, ニュースのテーマを指すような長い文字列を除外する, 異なるドメインへのリンクを除去する, などがある. 表 3.1 に属性と見なせないアンカーテキストを列記する.

表 3.1: 属性と見なせないアンカーテキスト一覧

こちら	ここ	戻る	トップ	ジャンプ

これらについては適時ストップワードとして設定し, 異表記同義語の同定では考慮しない.

また, 一般的でない属性があり. 例えば, 「東京サテライトキャンパス」, 「新キャンパス構想」などがあげられる. これらは, 特殊な表現のアンカーテキストは出現頻度が小さいという予想の元, 頻度の高いアンカーテキストに絞ることで対処する.

また, 異表記同義語が存在するため様々な表現が抽出されてしまう. 例えば学校案内に関するページにおいて, サイト A では「学校案内」となっているが, サイト B では「本校

について」, サイト C においては「こちら」という表記でアクセス者を誘導している場合がある. これについては後述する異表記同義語の関係を抽出することで対応する.

3.1.4 アンカーテキストの異表記同義語

アンカーテキストにおける異表記同義語関係の例を図 3.4 に示す. Site A のトップページには「入学情報」「学生生活」「教育・研究組織」という文字列でそれらに関するページへのリンクが設定されている. 一方, Site B のトップページでは「入学案内」「キャンパスライフ」「大学プロフィール」という文字を使用しそれらに関するページへリンクが張られている. ここでは「入学情報」と「入学案内」, 「学生生活」と「キャンパスライフ」が異表記同義語関係であることを示している.

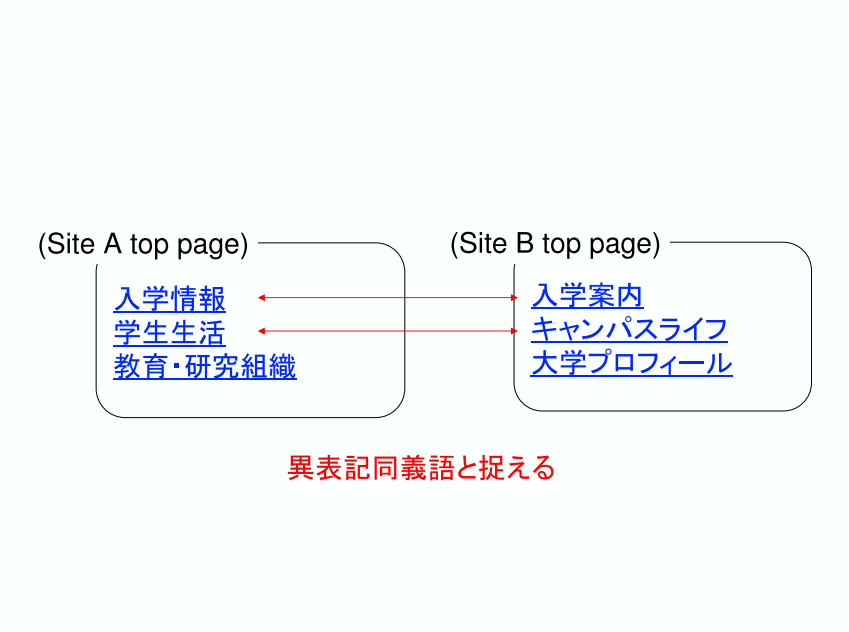

図 3.4: 異表記同義語関係

本研究では, ある二つのアンカーテキストを対象としたとき, 互いに同じもしくは似たような情報が含まれる参照先ページへリンクされている場合, 異なる表現ではあるが同じような意味と捉え異表記同義という言葉を使用している. 異表記同義の関係にある語を一見しただけでは同じような意味になると想像がつかない語も含まれるため, 一般的に使われている「類義語」や「同義語」の定義とは若干異なることを踏まえ, 本研究では異表記同義という表現を使用する.

3.2 異表記同義語関係の抽出の流れ

以下に, 異表記同義語関係の抽出の流れを示す.

1. トップページ (主に index.html, index.htm, top.html などがファイル名となっているページ) 中のアンカータグの箇所からアンカーテキスト, リンク先のタイトル文字列, URL を抽出する. (主にテキストファイルへのアンカーのみ対象とする)
2. URL をもとにいくつかの階層 (リンク深さ, ディレクトリ階層ではない) 分か繰り返す. (階層数については実験や他の論文などを通し最適な値や, 動的に変更することを考える)
3. 同クラスの他のいくつかのサイトにて 1, 2 の操作を行う.
4. 抽出結果にある基準を設けクラスに対する属性に重みを付ける.
5. 重み付けの結果をソートし, どの程度の重みでクラスに適した属性情報が抽出されるか検討する.

以下, 各 Step の詳細を述べる.

3.2.1 Step1:リンク情報の抽出と適切なアンカーテキストの抽出

クラス毎にサイトの HTML 文書を収集し, リンク情報を抽出する方法を述べる. ここでリンク情報とは, 文書内にアンカーテキストによって記述されているアンカーテキストと URL のペアを指す. まず, 調査対象のサイトのトップページをダウンロードし, アンカータグでしめされる他の文書への URL と誘導用の文字列 (アンカーテキスト) を抽出する. 抽出語の例を表 3.2 に示す.

「–」以降に, アンカータグに記述されている参照先の URL, アンカーテキスト, アンカーテキストを品詞分解しその最後の品詞, 品詞数を示している. 品詞は一般的ではない属性の排除, 品詞数は属性と見なせないアンカーテキストの排除として利用する. (現時点では排除の基準が定まらなかったため未実装)

表 3.2: アンカーテキストの抽出例

DownloadFile: www.jaist.ac.jp/index-j2.shtml
UrlPath: www.jaist.ac.jp/
Title: 北陸先端科学技術大学院大学 [JAIST]
–
www.jaist.ac.jp/index.html, 北陸先端科学技術大学院大学 [JAIST], 名詞, 4
www.jaist.ac.jp/index-e.html, English page, 名詞, 2
www.google.co.jp, google, 名詞, 1
www.jaist.ac.jp/index.html, ホーム, 名詞, 1
www.jaist.ac.jp/for_student.html, 受験生の方へ, 助詞, 4
www.jaist.ac.jp/for_society.html, 一般・社会人の方へ, 助詞, 7
www.jaist.ac.jp/for_company.html, 企業の方へ, 助詞, 4
www2.jaist.ac.jp/gakunai/gakunai.html, 学内情報, 名詞, 2
www.jaist.ac.jp/to_outline.html, 大学案内, 名詞, 2
www.jaist.ac.jp/to_establishment.html, 教育・研究組織, 名詞, 4
www.jaist.ac.jp/to_admission.html, 入学案内, 名詞, 2
www.jaist.ac.jp/to_campuslife.html, 学生生活, 名詞, 2
www.jaist.ac.jp/to_cooperation.html, 交流・連携, 名詞, 3
www.jaist.ac.jp/ks/index.html, 知識科学研究科, 名詞, 4
www.jaist.ac.jp/is/index-jp.html, 情報科学研究科, 名詞, 4
www.jaist.ac.jp/ms/index.html, マテリアルサイエンス, 名詞, 1
www.jaist.ac.jp/satellite/sate/newcampus, 東京サテライトキャンパス, 名詞, 3
www.jaist.ac.jp/ kouhou/General_info/access/access.html, 交通案内, 名詞, 2

抽出結果から, 属性と見なせないアンカーテキストを以下の基準を設け除外する.

1. 文字列長の長いアンカーテキスト
2. 異なるドメインを参照しているアンカーテキスト
3. 一般的な属性とならないアンカーテキストの削除

1 はニュースのタイトルなど助詞が多く含まれているアンカーテキストを指す. 2 は複数のドメインにわたって構築されているサイトについては本研究では考慮しないこととする.(表 3.2 では www.google.co.jp が相当する) 3 は表 3.1 で示したアンカーテキストなどを除外することを意味する.

3.2.2 Step2:索引語ベクトルの生成と文書の分類

本研究ではアンカーテキスト間の類似性を求めるために文書の本文から索引語ベクトルを生成し, k-means 法によるクラスタリングで各ベクトルの類似性を求め異表記同義語関係を求めている. 以下に Step2.1 で索引語ベクトルの生成法, Step2.2 で文書の分類方法について述べる.

Step2.1:索引語ベクトルの生成

Step1 で収集した HTML 文書群で出現頻度の高いアンカーテキストに注目し頻度や出現箇所により重み付けを行う. アンカーテキストが参照している先の文書の索引語ベクトルを生成する手順を以下に示す.

索引語の抽出

本研究では HTML 文書中の本文とアンカーで表現される URL から索引語を抽出している. まずは本文から索引語を抽出する手法について説明する. アンカーテキストが参照している文書から, HTML や各種スクリプトなどのタグを除去する. ブラウザを通して人間の視覚域に現れる文字列のみになった文章 (本稿では「本文」と表現する) にする. 抽出した本文を形態素解析器 Mecab[8] により品詞分解し, キーワード自動抽出システム Termextract[21] を使用し出現頻度と連接頻度をもとにした複合語を求め索引語 (複合語のみならず単語も含む) とする. なお, 索引語は全て日本語に限定した.

次に URL から索引語を抽出する方法について述べる.

一般に URL は「www.jaist.ac.jp/gakusei/guidance/kishukusha.html」「www.jaist.ac.jp/is/jpn/about/access.html」などトップドメインと参照ファイル (ここでは kisyukusha.html や access.html を指す) の間に階層を設け管理しやすい形になっている. その点を考慮し, 「gakusei」となっていれば「がくせい」と平仮名に変換し索引語とする. 本文からの索引語に「学生」があった場合などは Mecab による「読み」方情報から「ガクセイ」を抽出しマッチしていれば頻度に上積みする. マッチングの基準はエディットディスタンスを基準とした. ディレクトリ名は比較的短い文字列であると判断し比較対象の文字列のエディットディスタンスが 1 以下, もしくは 70%以上の一致性が見られたときはマッチングしているものとした.

文書毎に索引語の出現頻度を計算

求めた索引語がその文書に出現している頻度を求める. (Termextract の内部処理で出現頻度を計算させ独自のスコアを出力しているが, Termextract にはキーワード生成のみを行わせ, ここでは改めて出現頻度 (のべ数) を算出している)

ベクトル生成

本研究ではベクトル生成を単純な出現頻度と, 求めた出現頻度と全文書に対する出現文書数をもとにした $TF \cdot IDF$ での重み付けで行っている. 単に出現頻度のみである文書の索引語はその文書をどの程度特徴づけているのか不明であるためである. 例えば「学校案内」に関する文書があったとする. 本文中の用語に学校案内に関する用語「案内」「学校案内」などが含まれていれば出現頻度が高くなりその文書を特徴づけている用語として文書類似度で効果を発揮するであろう. しかし, 現在インターネット上の文書は単にテキストデータに限らず様々なマルチメディアを駆使し作成されているサイトが多く存在する. そのため「学校案内」に関するページにそれに関する文字列が含まれない場合が存在すると考えられる. 以上から本研究ではベクトル生成時に $TF \cdot IDF$ での重み付けを行い, 特定の少数の文書に出現する索引語に大きい重みを与える [12]. 式 3.1 に $TF \cdot IDF$ の計算式を示す. ここで, N は検索対象となる文書集合中の全文書数, $df(t)$ は索引語 t が出現する文書数である.

$$idf(t) = \log \frac{N}{df(t)} + 1 \tag{3.1}$$

なお出現頻度は, 本研究室の検索システムと YahooAPI[32] を利用した.

Step2.2:文書の分類

Step2.1 で生成した索引語ベクトルを元に k-means 法による文書分類を行う. 局所解に対しては何回か[1]のクラスタリングを行い, 目的関数 (セントロイドと割り当てられたサンプルの距離の総和) が最小となる結果を選択する.

3.2.3 Step3:異表記同義語関係にあるアンカーテキストの同定

異表記同義語関係にあるアンカーテキストの同定について述べる.

クラスに適当な属性の決定

まず, クラスに適当な属性を決める. 各クラスに対する属性の決定を, 以下の手順により行う.

1. サイトの URL をディレクトリ, ファイル名に分解
2. 単数形変換, 記号除去などで単語を抽出
3. 抽出された単語に対して, 他のサイトの単語の出現頻度を調査
4. 出現頻度により属性を決定

[1]実験では 25 回で行った。

各属性に適合するクラスタからの決定

抽出した各属性に対して最も適合するクラスタを決定する. (詳細は第 4 章を参照)

3.3 システム概要

提案手法の具体的な実装基準を述べる.

本研究で開発したサイト情報抽出システムは抽出処理と解析処理に大別される. 抽出処理は, 文書収集処理, タグ情報抽出処理を行う. 一方解析機能は, 抽出処理の出力を入力とし, 索引語抽出処理, 索引語ベクトル生成処理, クラスタリング処理, 異表記同義語関係を同定する処理を行う.

抽出機能が出力した索引語とタグ情報を入力とし, 索引語の出現頻度, 各種重み付けにより索引語ベクトルを生成し, クラスタリング手法の一つである k-means 法を用い異表記同義語の分類を行う. 以下, 図 3.5 に本システムの概要を示す.

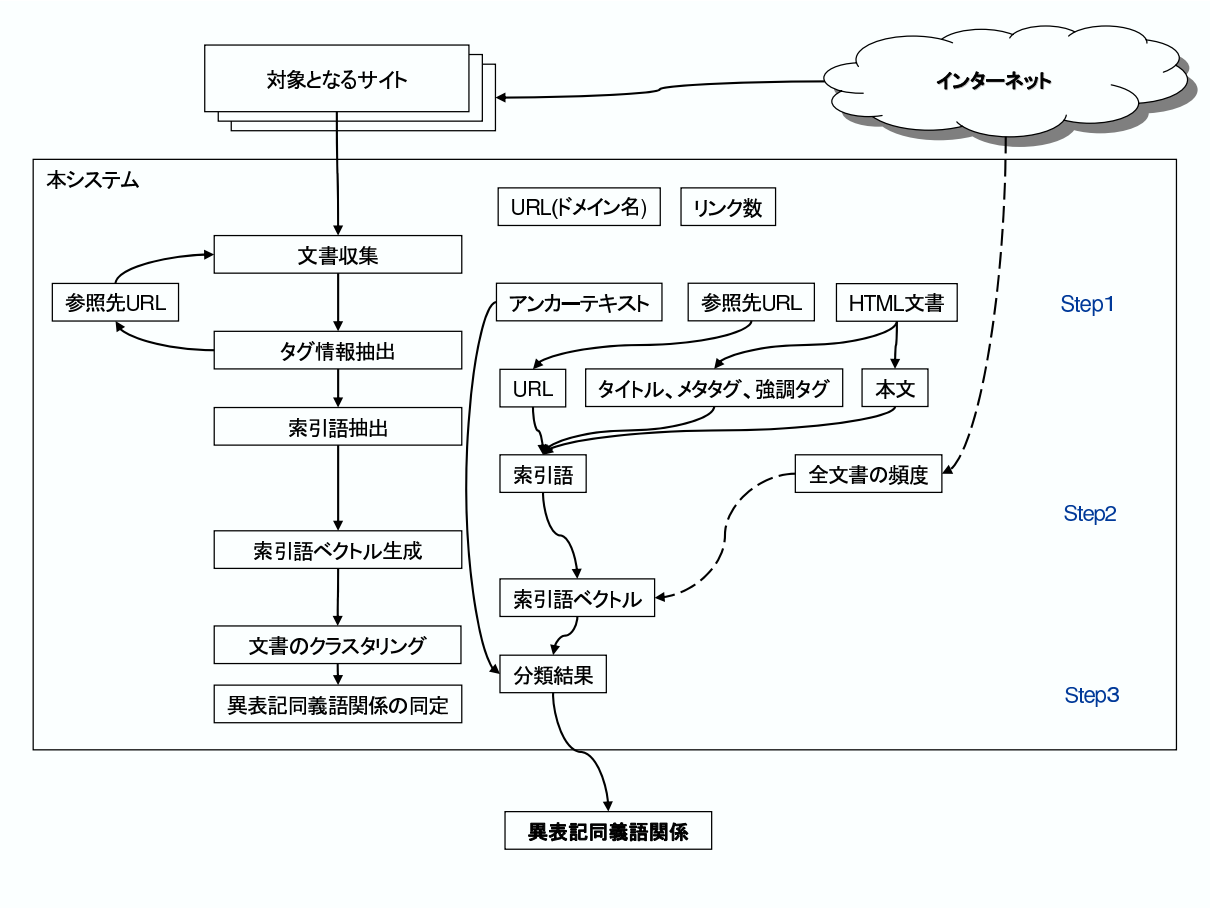

図 3.5: システム概要

3.3.1 文書収集処理

文書収集処理について述べる.

転送処理はリンク情報 (アンカーテキストと参照先 URL のペア) の抽出を行いながら HTML 文書 (HTML によって記述される文書) を収集する.

wget[19] を使用しインターネット上の web サーバから文書のダウンロードを行う. wget にはリンクの深さを指定できるオプションがあるが, 深さに関する情報が出力されないため, 本システムの転送処理機能で深さを捕捉できるようにしている.

また, 対象とした HTML 文書は, 拡張子が html, htm, cgi, asp, jsp のファイルとした. 拡張子なしファイル (例. 参照先 URL が www.sample.com/sample) も含めているが, wget によってダウンロードされたファイルが HTML によって記述されている文書だった場合は通常の処理を行い, 実体が異なる場合 (例. 上記例の実体が www.sample.com/sample/index.html であった場合) は転送対象ファイルの URL を実体の URL でリンク情報 (アンカーテキストと参照先 URL のペアを格納している情報) の更新を行う.

参照先 URL とダウンロードされるファイルが異なる場合 (参照先 URL がドメイン名のみ, ディレクトリ名のみの場合) はそのペアの情報を保持し後続の処理に継がせる. 以下に処理の様子を示す.

1. 参照先 URL が, 参照先 URL と実体 URL のペアとして保存されているか確認
2. 保存されていた場合, 実体 URL の文書で後続処理
3. 保存されていない場合, 以下の処理を行う
4. wget で参照先 URL を入力
5. wget のログで実際にダウンロードした URL(実体 URL) を確認
6. 入力した参照先 URL と実体 URL のペアを保存

3.3.2 タグ情報抽出処理

タグ情報抽出処理について述べる.

ダウンロードした HTML 文書のタグを解析しアンカーテキスト, アンカーテキストとペアの参照先 URL, メタタグ, 強調タグ (表 3.3 参照) を元にそれらの内容を抽出する.

本文, 参照先 URL は索引語ベクトル生成 (後述 3.2.2) で使用する. タイトルはその文書を一言で表現する特徴として索引語ベクトル生成時の重み付けに使用する. メタタグ (主に keyword タグ), 強調タグで示される内容も重み付けに使用する.

なお, フレーム対応文書 (frame タグあり) の場合, frame タグで参照されている URL から文書を抽出しフレームなし文書の操作で抽出した内容を結合し, その文書の情報とする. ただし, frame タグで参照された先の文書がフレーム対応文書ではない場合処理を行う. 表 3.6 にフレーム対応文書からの抽出例を示す.

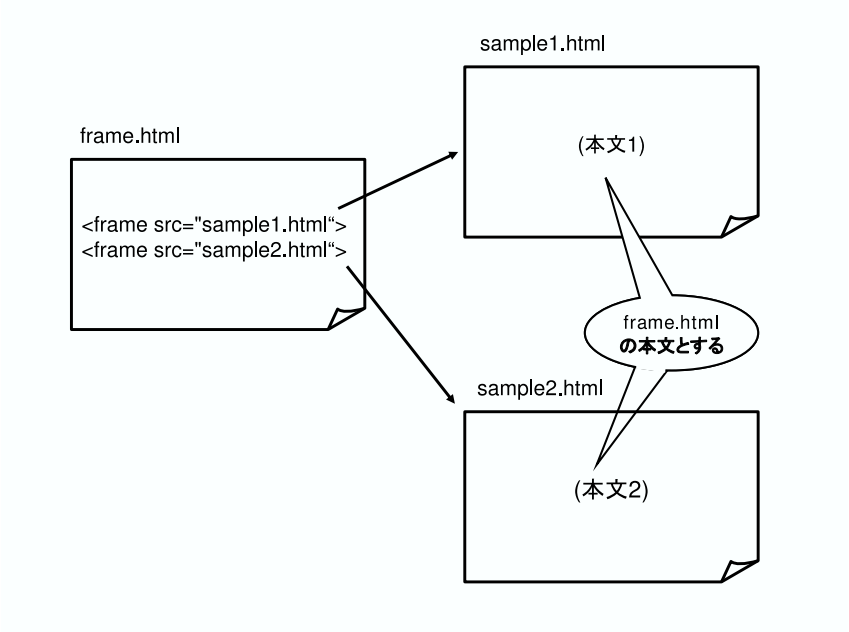

図 3.6: フレーム対応文書からの抽出例

3.3.3 索引語抽出処理

索引語抽出処理は索引語ベクトル生成に使用する索引語を抽出する処理である. 基本的抽出, URL 強調時の抽出, アンカーテキスト強調時の抽出に分け説明する.

基本的抽出

HTML 文書から抽出した本文のみのファイルを形態素解析器 Mecab に入力する. ここで本文とは body タグで囲まれ, 他のタグ情報を除去したブラウザで人が見える文字列を主とし, それに加え, タイトル, メタ情報を加えたものとしている.

形態素解析器によって出力された品詞分解結果をキーワード生成器 TermExtract に入力させ索引語を出力させる. なお, 索引語は日本語のみを対象としている. 以下に索引語抽出の例を示す.

(入力した文)
大学周辺には，教員などの職員が入居する職員宿舎と，主に学生が入居する学生寄宿舎があります。特に学生寄宿舎は，鉄筋コンクリート５階建ての建物８棟全てがキャンパス内にあり，大学と寄宿舎の一番近い建物同士だと，その距離は数十歩。中には１００歩足らずで自分の研究室に着いてしまうという人もいて，研究熱心な方には申し分ない立地条件です。
中身はというと，専攻分野，経歴などにとらわれず広く学生を受け入れるという本学の理念にふさわしく，単身室，夫婦室，家族室という充実のラインナップ。ご家族のいる社会人の方も安心して研究ができます。大学のある丘のふもとには保育園と小学校もあります。
この寄宿舎には，一般の学生はもとより，留学生，そして本学が海外から受入れた外国人研究員も入居しています。大学内だけでなく，普段のご近所づきあいでも国境を越えたインターナショナルな雰囲気が楽しめます。学生寄宿舎について興味を持たれた方は

→

(出力された用語)
学生, 学生寄宿舎, 大学, 寄宿舎, 本学, 入居, 留学生, 研究, 大学内, 職員宿舎, 研究室, 外国人研究員, 大学周辺, 小学校, 研究熱心, 建物, 職員, 家族室, 教員, 家族, 建物同士, 受入, 単身室, 社会人, 一番近, 中身, 国境, 自分, 階建, 夫婦室, 数十歩, 専攻分野, 近所, 立地条件, 海外, 歩足, 保育園, 雰囲気, 鉄筋コンクリート, キャンパス内, 安心, 一般, 理念, 経歴, 普段, 距離, 棟全, 充実, 興味

URL 強調時

URL 強調時の索引語生成処理について述べる.

URL はアルファベットで記述されているという仮定の下, カタカナ変換 (suikyo[7]), 日本 (gene95[25]) 語変換, 単数形変換 (morph[30]), 同義語データベース (WordNet[20]) を使用し強調を行う.

アルゴリズムを以下に示す.

1. URL からドメイン名を除去
2. URL を “/” で分解
    - 記号 (“_” も含む[2]) が入っていればさらに分解
3. 分解した単語を morph を使用し単数形に変換
4. 各単語について, WordNet で抽出した同義語を結合
5. 各単語を gene95 を使用し日本語に変換
    - 変換できなかった場合, suikyo でカタカナ語に変換

[2]Ruby では正規表現 “/\W/” では “_ (アンダーバー)” にマッチしないので注意

- さらに変換できなかった場合, 無視

6. 各単語を 3.3.4 のベクトル生成で設定した任意の倍率で重みを強調
7. 強調する単語が文書に無かった場合, 各単語中にカタカナ語があればそれを索引語に追加し重み付け

アンカーテキスト強調時

アンカーテキスト強調時の索引語生成処理について述べる.

抽出済みの索引語がアンカーテキストに含まれていれば, その索引語の重みを任意の倍率に従って重み付ける. 含まれていなければ, 処理を行わない.

本研究では上記実装にて実験を行ったが, アンカーテキスト強調時にはアンカーテキストの文字列をキーワード生成器にかけ, 抽出されたキーワードを索引語として追加する処理も考慮した. しかし, ある文書に関連づけされているアンカーテキストは複数抽出される. Step1 の処理でリンクの深さを変更することで, 一つの文書に対して多くのアンカーテキストが関連づけされた場合, アンカーテキストから索引語を生成するとしたら, 複数のアンカーテキストを全て含めその文書に対する索引語を生成することになるであろう. となると結局, アンカーテキストから索引語を生成し上記基本的抽出での索引語と混合させると, リンク数に依存したベクトルを生成することになる. よって, 今回の実験ではアンカーテキストからは索引語候補を抽出せず, 既存索引語とのマッチング処理のみを行う.

3.3.4 索引語ベクトル生成処理

索引語ベクトル生成処理についてを述べる.

タグ解析による文書の特徴要素 (タイトルとメタタグ) と索引語を元に各文書の索引語ベクトルを生成する. 強調タグは表 3.3 のタグを用いている.

表 3.3: 強調タグ一覧

強調タグ	意味
em	文字を強調する
strong	さらに強調する
h1	見出し文字 (大見出しなどに利用)
h2	見出し文字 (中見出しなどに利用)
h3	見出し文字 (小見出しなに利用ど)
h4	見出し文字
h5	見出し文字
h6	見出し文字
th	見出し文字
big	大きめにする
blink	文字を点滅させる (ブラウザ依存)
marqueeb	文字をスクロールさせる (ブラウザ依存)
li	リスト
i	斜体文字にする
u	文字の下に線を引く
tt	等幅フォントを使用
center	囲まれた内容を中央に表示

$TF \cdot IDF$ は, 本研究室にて構築されている検索システムによる文書数, YahooAPI[32] による文書数を使用し 3.2.2 で示した方法により算出する.

全文書数について, 本研究室の検索システムの総数はわかっているが, YahooAPI の場合については公表されていないため以下の方法により総数を求めた. まず, 本システムでは日本語で記述されている文書を対象とし, 索引語を日本語で抽出しているため, 検索対象の文書集合を日本語で記述されている集合に限定した. ここで日本語で記述されている全文書数の特定として平仮名の一文字を検索クエリーとしヒット件数を算出した. 以下, 3.4 に上位 10 のヒット件数を示す.

表 3.4: 上位 10 のヒット件数 (2007 年 2 月時点)

順位	ひらがな	ヒット件数
1	の	1,800,000,000
2	は	1,550,000,000
3	を	1,510,000,000
4	に	1,500,000,000
5	と	1,350,000,000
6	お	1,230,000,000
7	も	1,030,000,000
8	な	996,000,000
9	へ	990,000,000
10	か	874,000,000

上記結果により平仮名「の」のヒット件数 (1,800,000,000 [3]) を全文書のヒット件数とし, 各索引語に対して「の」を含めた検索を行う. (例. 索引語が「研究室」の場合, 検索対象語を「の 研究室」とする) 図 3.7 に全文書数算出のイメージを示す.

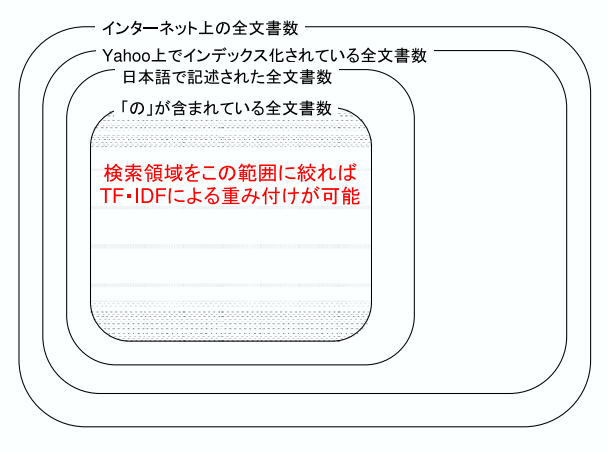

図 3.7: 全文書数算出のイメージ

3.3.5 文書のクラスタリング

文書のクラスタリングについて述べる. 索引語ベクトルを元にクラスタリングを行い各文書の類似関係を求める.

クラスタリング手法については比較的単純なアルゴリズムである k-means 法を採用した. k-means 法のアルゴリズムを以下に示す.

[3]アルファベットで検索したところトップは “a” の 8,020,000,000 件であった

1. 各データ $x_i(i = 1 \cdots n)$ に対してランダムにクラスタを割り振る.
2. 割り振ったデータをもとに各クラスタの中心 $V_j(j = 1 \cdots)$ を計算する. 計算は通常通り当てられたデータの各要素の平均 (重心) を使用する.
3. 各 x_i と各 V_j との距離を求め, x_i を最も近い中心のクラスタに割り当て直す.
4. 上記の処理で全ての x_i のクラスタの割り当てが変化しなかった場合は処理を終了する. それ以外の場合は新しく割り振られたクラスタから V_j を再計算して上記の処理を繰り返す.

ただし, クラスタリング結果はランダムに割り振ったクラスタの初期値に大きく依存することが知られているため, 局所的最適解にすぎない場合が考えられる. 本研究では 10 回以上クラスタリングを行い, その中で最も目的関数が小さかった結果を大域最適に近い解として出力させている.

類似度測定手法

ベクトル間の距離を測定する方法として, コサイン尺度とユークリッド距離で実装し比較実験を行った.

ここでコサイン尺度とは, ベクトル間の類似度を求める手法として文書検索でよく用いられているものであり, $\boldsymbol{q}$ を検索質問ベクトル, 各文書ベクトルを $\boldsymbol{d_j}$ とすると次の式 3.2 となることが知られている.

$$cos(\boldsymbol{d_j}, \boldsymbol{q}) = \frac{\boldsymbol{d_j} \cdot \boldsymbol{q}}{||\boldsymbol{d_j}||||\boldsymbol{q}||} \tag{3.2}$$

また, n 個の実数の組全体の集合 $\boldsymbol{R^n}$ の二点 $\boldsymbol{a} = (a_1, a_2, ..., a_n)$, $\boldsymbol{b} = (b_1, b_2, ..., b_n)$ を考えるとユークリッド距離 $d(\boldsymbol{a},\boldsymbol{b})$ は式 3.3 のようになる.

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{3.3}$$

比較実験を図 3.8, 3.9 に示す.

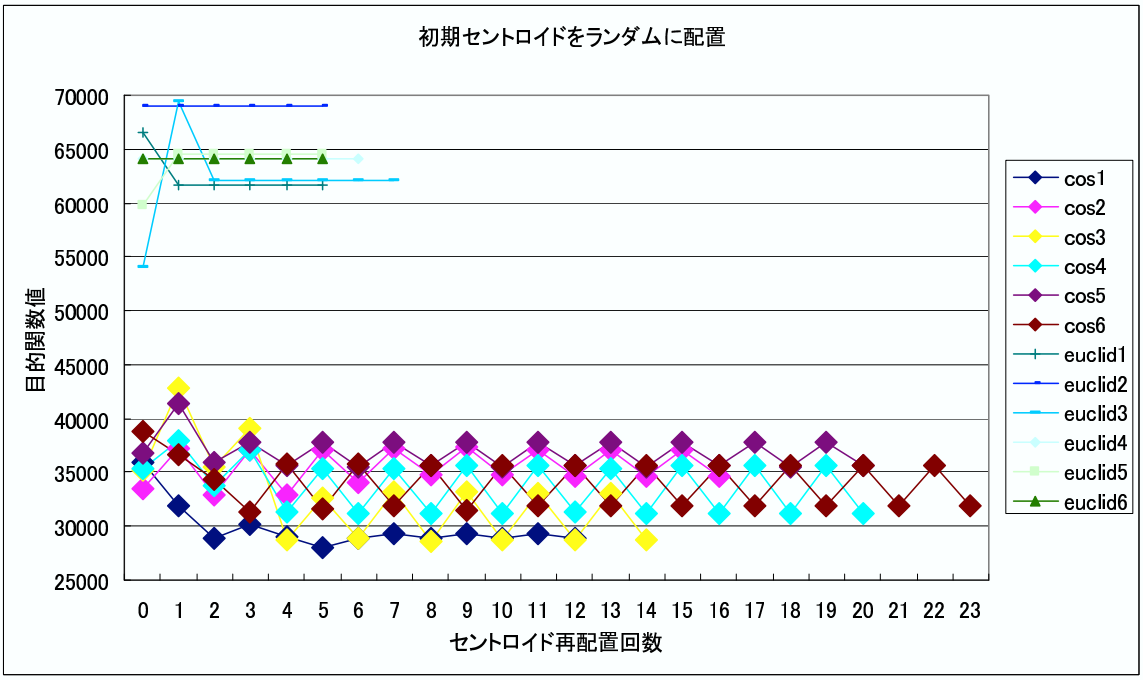

図 3.8: 初期セントロイドをランダムに決定

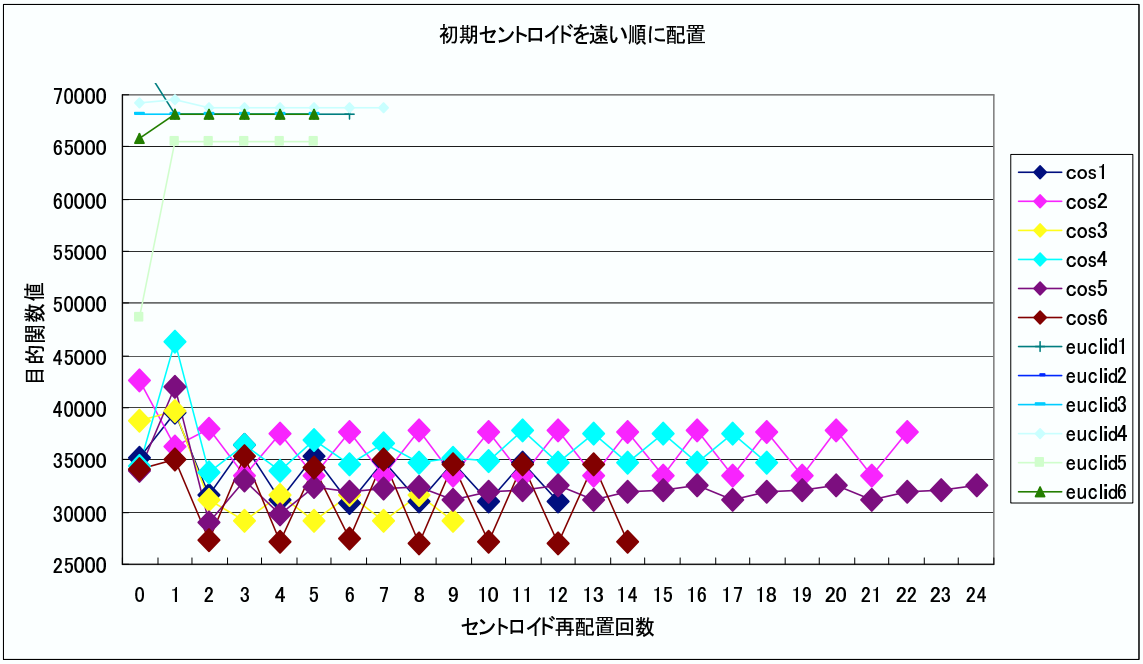

図 3.9: 初期セントロイドを遠い順に決定

図 3.8, 3.9 での目的関数値はセントロイドとそのセントロイドに割り当てられた全サンプルの距離 (ユークリッド距離で測定) の総和であり, 何回かの試行回数により目的関数値の変動の様子である. 図 3.8 では初期のセントロイドをランダムに配置しクラスタリングを行った結果を示している. 距離を cos 尺度で計り最も近いセントロイドへ割り当てた場合では目的関数の値が収束せず, ユークリッド距離での割り当てでは 3 回ほどで収束している. 図 3.9 では初期のセントロイドをランダムではなく, 各サンプルの距離を求め最も離れているサンプル値を初期のセントロイドとして割り当てクラスタリングさせたた結果を示した. これはセントロイドをランダムに割り当てたものと比べ, よりばらついた位置にセントロイドを割り当てることで局所解の出現を抑え最適解が導かれるという予想の元実験を行ったが, 目的関数の値を観察したところ目立った改善はされなかった.

3.3.6 異表記同義語関係の同定

クラスタリング結果からアンカーテキストにおける異表記同義語関係を導く.

属性項目抽出

URL に対して以下の操作を施し属性候補を抽出する.

1. ドメイン名を除外
2. ディレクトリ名とファイル名 (拡張子を除外) に分解
3. 記号[4]もしくは_(アンダーバー) で分解
4. 表 3.5 に示すストップワードで除去
5. 複数形の単語を単数形へ変換[5]
6. 上記操作を各サイトで行い, 他のサイトで抽出した属性候補と比較
7. 多数のサイトで使用されている属性名であったとき, 属性名として抽出

表 3.5: 属性項目のストップワード

index	top	main	home	pub	default
image	img	frame	cgi-bin	html	bin
pdf	www	cgi	(数字のみ)	~ (先頭がチルダ)	%7E
(先頭が “7E”)	(先頭が “?”)	2 文字以下			

属性候補を他のサイトと比較する際, 本実験では他の 1 サイト以上で同じ属性候補名が抽出されていればそのクラスの属性名であると判断した.

これらの属性候補で, 初期クラスタを導く.

英単語の場合, 英和辞書を用い日本語に変換後, 索引語候補に出現した語で初期クラスタを生成する. アルファベットで記述された日本語 (“bosyu”, “gairai”, “gaiyou” など) の場合, 索引語候補を検索[6]し, ヒットした索引語候補から初期クラスタを生成する.

[4]正規表現 “/\W/” を使用
[5]morph[30] を使用
[6]ローマ字のまま日本語を検索可能な Migemo[11] を使用

第4章　実験

本章では, 実装したシステムを用い各種パラメータ (4.2.1 参照) を強調し重み付けを行った索引語ベクトルに対してクラスタリングを行う. クラスタリング結果を, 人手による正解データ (以下, 単に正解データと呼ぶ) と比較し F 値で評価を行う.

4.1　目的

各種パラメータを強調した場合と強調していない場合の比較を行い, 各種パラメータによる強調操作の有効性を検証する. 評価値として F 値を使用する.

4.2　方法

以下に, 実験手順, 評価方法について述べる.

4.2.1　実験手順

以下に, 実験手順を示す.

1. 文書収集
2. パラメータ設定
3. クラスタリング
4. 正解データとの比較
5. 検証

パラメータは以下の項目を対象とする.

- クラスタ数
- アンカーテキストによる強調倍率
- URL による強調倍率
- タイトルによる強調倍率

- メタ情報による強調倍率
- 強調タグによる強調倍率

また, TF と $TF \cdot IDF$ の比較も行う.

4.2.2 評価方法

クラスタリング結果と正解データを比較することで評価を行う. 再現率 $R(recall)$ と適合率 P(精度, $precision$) から F 値を算出し, 評価値とした.
算出方法を式 4.2, 4.1, 4.3 に示す.

$$R = \frac{\text{クラスタに含まれる対象の属性語に属する } URL \text{ 数}}{\text{対象の属性語に属する } URL \text{ 数}} \tag{4.1}$$

$$P = \frac{\text{クラスタに含まれる対象の属性語に属する } URL \text{ 数}}{\text{クラスタに含まれる全 } URL \text{ 数}} \tag{4.2}$$

$$F = \frac{2 * P * R}{P + R} \tag{4.3}$$

4.3 準備

本実験は以下の条件でを行った.

表 4.1: 実験条件

	クラス	サイト数	全URL 数	入力クラスタ数
条件 1	研究室	5	68	5, 11, 18
条件 2	病院	10	195	5, 12, 20, 28, 35

4.4 結果

以下に, 予備実験と各種パラメータの強調効果の結果を述べる.

4.4.1 予備実験

予備実験ではクラスタリング毎のF 値の散らばり (ばらつき) と, 属性候補の自動抽出を行った. 以下に結果を述べる.

クラスタリング毎の F 値の散らばり

まず, クラスタリング結果と人手の正解データの比較として算出した F 値について, クラスタリング毎に散らばりが見られたため, その結果を図 4.1, 図 4.2 に示す.

実験は条件 1(5lab), 条件 2(10byouin) にて行った.

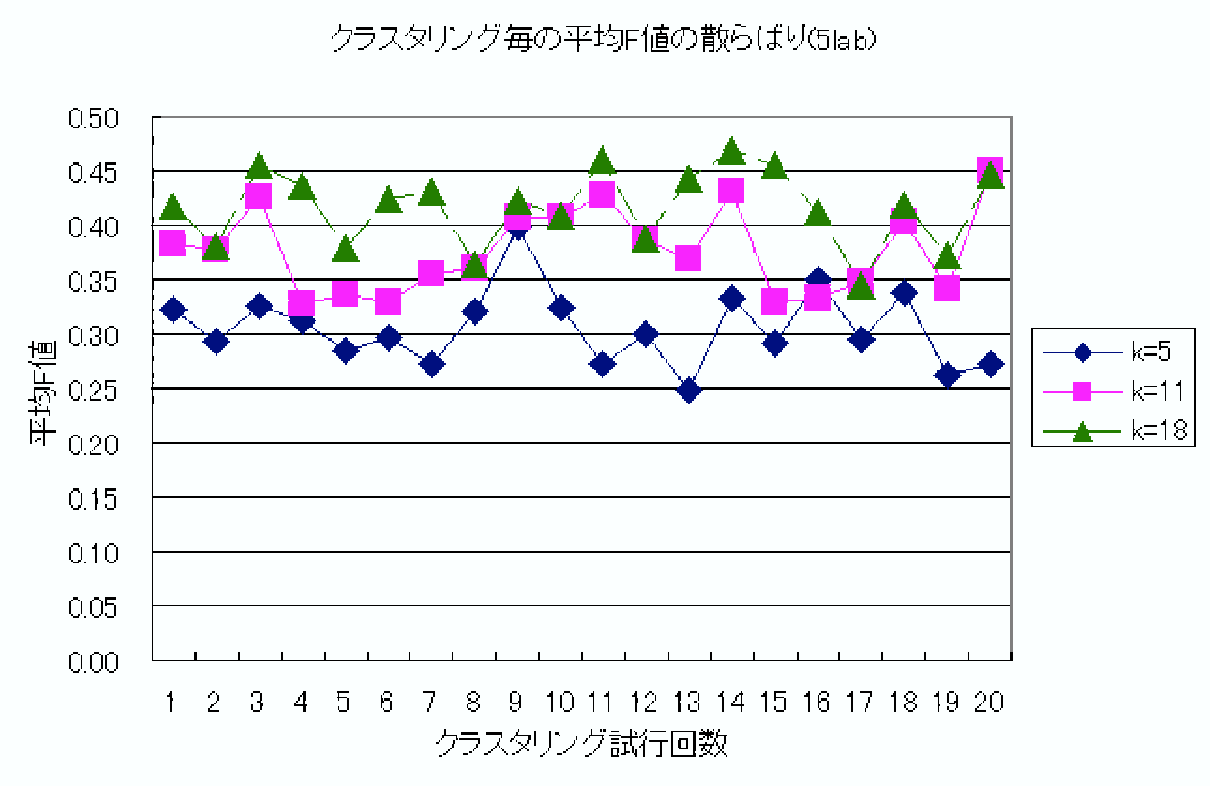

図 4.1: クラスタリング毎の平均 F 値の散らばり (5byouin))

図 4.1 はクラスタ数を 5, 11, 18 についてクラスタリングを 20 回試行した F 値の平均を示している. k=5 で 0.15, k=11, 18 で 0.12 の範囲で F 値に散らばりが見られる.

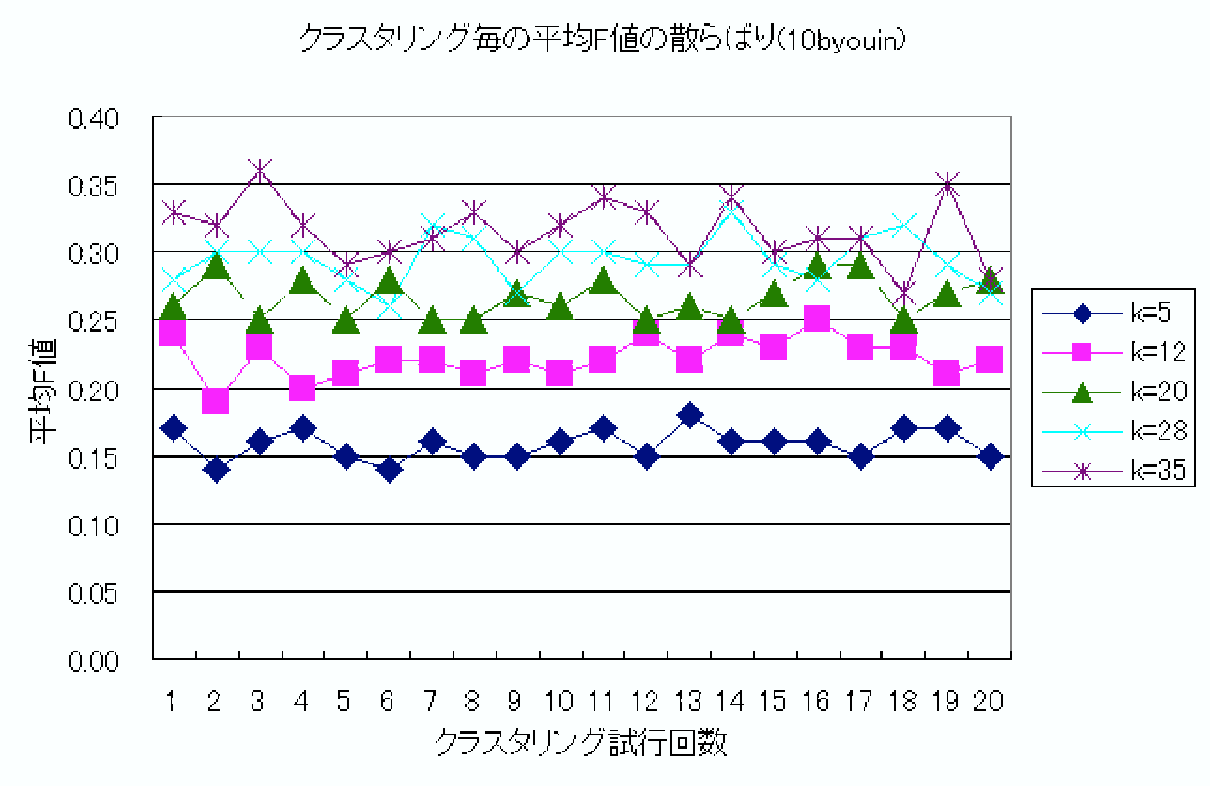

図 4.2: クラスタリング毎の F 値の散らばり (10byouin)

図 4.2 はクラスタ数を 5, 12, 20, 28, 35 についてクラスタリングを 20 回試行した F 値の平均を示している. k=5 で 0.04, k=12 で 0.06, k=20 で 0.04, k=28 で 0.07, k=35 で 0.09 の範囲で F 値に散らばりが見られる.

属性候補の自動抽出

属性候補の抽出結果を示す.

表 4.2: 自動抽出された属性名 (条件 1, 研究室 5 サイト)

access	member	paper	publication	research

表 4.3: 自動抽出された属性名 (条件 2, 病院 10 サイト)

access	bosyu	gairai	gaiyou	info	kensin
map	nintei	nyuin	renkei	rinen	sisetu

表 4.4: 自動抽出された属性名 (ラーメン屋 10 サイト)

info	link	map	menu	new	recruit	shop

表 4.5: 自動抽出された属性名 (ホテル 10 サイト)

access	bath	yoyaku	contact	enkai	event	facility	roten
form	group	history	honkan	info	kankou	lady	sitemap
link	map	onsen	plan	privacy	restaurant	room	wed

また, 自治体 10 サイトの自動抽出された属性名を付録に添付する.

4.4.2 異表記同義語抽出

以下に, 2 つのクラス (研究室, 病院) で, それぞれ 5 つのサイトをサンプルとし実験を行い異表記同義語を抽出した結果を示す.

研究室クラスでは, 「メンバー」「リサーチ」「アクセス」に着目し, 主観で選んだ正解データと実験結果を比較し, 属性毎にクラスタにまとまっているか確認したところ, 5 つの属性語が最小 2 つのクラスタにまとまった. 結果, 「Member」「構成員一覧」「メンバー」が異表記同義語関係として抽出された. 一方, 病院クラスでは「概要」「交通アクセス」「入院案内」に着目し実験を行ったところ, 5 つの属性語が最小 1 つのクラスタにまとまった. 結果, 「受診と入院の案内」「入院のご案内」「入院案内」「入院案内」「入院案内」が異表記同義語関係として抽出された.

研究室クラスのクラスタリング結果を表 4.6 に示す.

表 4.6: クラスタリング結果 (5lab)

anchor	URL power(1)			URL power(100)		
power	member	research	access	member	research	access
10	3/5	4/5	3/4	*2/5*	3/5	2/4
20	*2/5*	3/5	2/4	*2/5*	3/5	2/4
40	3/5	3/5	3/4	2/5	3/5	2/4
60	4/5	*2/5*	3/4	3/5	3/5	2/4
80	4/5	*2/5*	3/4	*2/5*	3/5	2/4
100	4/5	3/5	3/4	2/5	3/5	2/4
150	4/5	3/5	3/4	2/5	4/5	2/4
200	4/5	4/5	3/4	3/5	3/5	2/4

URL による強調を 1 と 100 の倍率, それぞれでアンカーテキストによる強調を 10 から 200 の倍率でクラスタリングさせている. 値が小さいものほど 5 つの対象物の属性語がまとまっていることを示しており, 属性「リサーチ (research)」の最小は URL 強調が 1 倍で, アンカーテキスト強調が 60, 80 倍の時に「2/5」とまとまっている. なお,「アクセス (access)」の母数が 4 になっているのは, 1 つのサイトで主観で属性「アクセス」に相当する文書が見つからなかったためである.

図 4.3 に属性「リサーチ」が 2 つのクラスタにまとまった様子を示す.

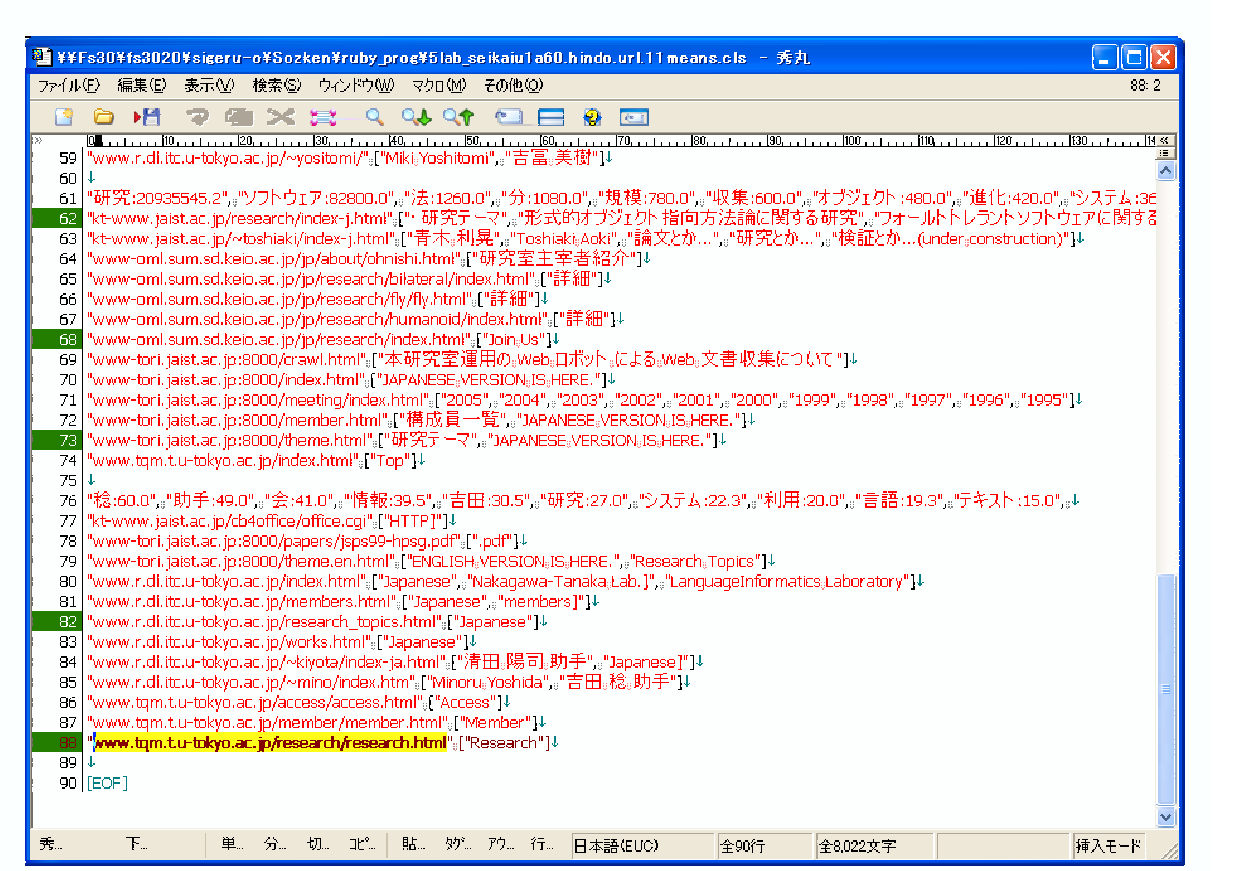

図 4.3: 最適クラスタリング結果 (5lab, research)

改行毎に 1 つのクラスタに相当する．URL の後の「[ ]」で囲まれている箇所が, 左の URL のペアとなっているアンカーテキストである．左側の数字がマーキングされている文書が主観で属性「リサーチ」に含まれると判断したものである．

上のクラスタに注目すると, 索引語「研究」に重みが大きく設定されており,4 つのサイトの文書が存在する．この結果から, 個人名や数字だけのものや「Top」「詳細」など属性とはなりにくいアンカーテキストを除去するとこのクラスタから抽出される異表記同義語関係は表 4.7 になる．

表 4.7: 抽出された異表記同義語関係 1

・研究テーマ	論文とか...	検証とか...	研究とか...
研究者主催者紹介	Join Us	構成員一覧	研究テーマ

これらの関係だけに注目するとそれぞれの語に対して, 「研究」という単語が 8 つ中 4

つに含まれており,「研究」という属性に割り当てられるべき異表記同義語関係ということができる.

病院クラスのクラスタリング結果を表 4.8 に示す. 表 4.8 は URL による強調を 1 倍にし, アンカーテキストによる強調を 10 から 200 倍の範囲で行っている. URL による強調を 100 倍にした結果を表 4.9 に示す. 正解データの属性語が最小にまとまったクラスタ数は 1 という結果が出ているが, 一方で属性「概要」に注目するとアンカーテキストによる倍率によりまとまり数が 3 から 5 と変動している. 属性「地図」に関しては 4 で強調による変動が見られなかった.

表 4.8: クラスタリング結果 (5byouin1)

	URL power(1)				
anchor power	gaiyou	map(access)	nyuin	sisetu	gairai
10	5/5	4/5	*1/5*	3/3	2/3
20	3/5	4/5	*1/5*	3/3	2/3
40	4/5	4/5	2/5	3/3	2/3
60	4/5	4/5	2/5	3/3	2/3
80	4/5	4/5	2/5	2/3	2/3
100	5/5	4/5	2/5	3/3	2/3
150	5/5	4/5	2/5	3/3	2/3
200	3/5	4/5	2/5	3/3	2/3

表 4.9: クラスタリング結果 (5byouin2)

	URL power(100)				
anchor power	gaiyou	map(access)	nyuin	sisetu	gairai
10	4/5	3/5	2/5	3/3	2/3
20	3/5	2/5	2/5	3/3	2/3
40	5/5	3/5	2/5	3/3	2/3
60	4/5	3/5	2/5	3/3	2/3
80	3/5	2/5	2/5	3/3	2/3
100	5/5	3/5	2/5	3/3	2/3
150	4/5	3/5	2/5	3/3	2/3
200	3/5	2/5	2/5	3/3	2/3

図 4.4 に属性「入院案内」が 1 つのクラスタにまとまった様子を示す.

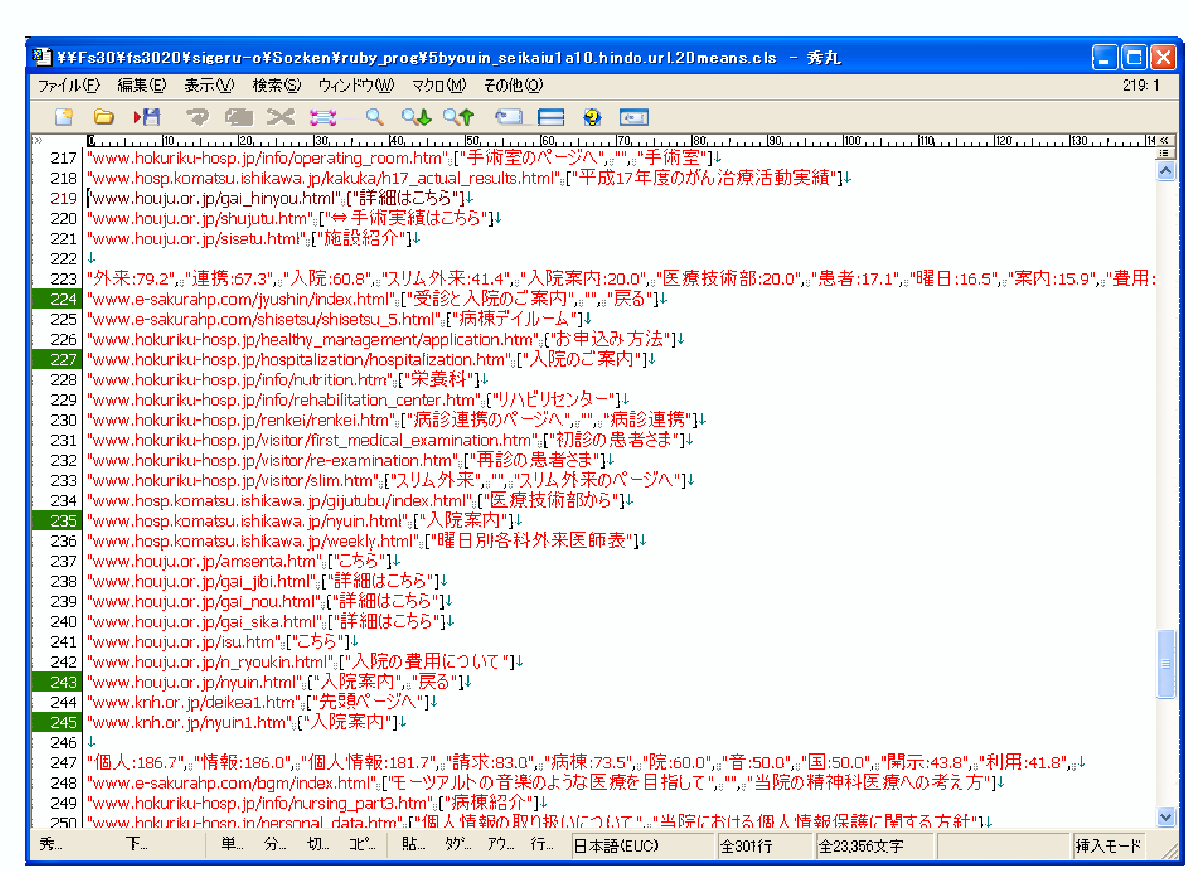

図 4.4: クラスタリング結果 1

このクラスタリング結果から, 索引語「外来 (頻度:79.2)」「連携 (頻度:67.3)」「入院 (頻度 60.8)」に対し比較的重みがついている. 理想的には索引語「入院」に対する重みが強いクラスタであれば,「入院」に特化したクラスタと判断できるが, この例では属性「外来案内」に相当する文書も割り当てられてしまうことが想像される.

割り当てられたアンカーテキストをもとに同定した異表記同義語関係を表 4.10 に示す.

表 4.10: 抽出された異表記同義語関係 2

受診と入院案内	病棟デイルーム	お申し込み方法	入院のご案内
栄養科	リハビリセンター	病診連携	スリム外来
医療技術部から	入院案内	曜日別各科外来医師表	入院費用について
入院案内 (houju.or.jp)	入院案内 (knh.or.jp)	初診の患者さま	再診の患者さま

この結果では多くの語に共通してみられるのは単語「入院」であると判断できる. しか

し, 16 の語の中で 7 つのみであり, かつ属性「外来案内」と同じクラスタに分類されているため, クラスタは 1 つにまとまっているが有効であるとはいえない.

4.4.3 各種パラメータの強調効果

以下に, 条件 1(研究室 5 サイト), 条件 2(病院 10 サイト) での実験結果を示す.

研究室 5 サイトでの実験

次に条件 1 にて行った, 各種パラメータの比較実験を示す. 表 4.11 に $TF \cdot IDF$ 重み付けによる比較 (全属性の F 値の平均と最大値) 結果を示す.

表 4.11: $TF \cdot IDF$ 重み付けによる効果 (全属性の F 値の平均と最大値)

	k5		k11		k18	
	平均値	最大値	平均値	最大値	平均値	最大値
TF	0.37	0.47	0.44	0.54	0.46	0.59
TFIDF1	0.34	0.41	0.42	0.48	0.43	0.51
TFIDF2	0.38	0.46	0.43	0.48	0.44	0.54

ここで, 表中 TF は索引語の文書中の頻度, TFIDF1 は本研究室の検索システム, TFIDF2 は YahooAPI[32] を使用した全文書の出現頻度である. 平均値に注目すると, TF, TFIDF の差は 0.02〜0.03 となり TFIDF による効果が見られたとはいいがたい. また, 入力クラスタ数 (k) と F 値が比例し増加している.

表 4.12, 4.13 にアンカーテキストの重み付けによる結果を示す.

表 4.12: アンカーテキストの重み付けの効果 (k=18 での全属性の平均 F 値の平均)

(倍率)

	1	5	10	50	100
TF	0.46	0.47	0.47	0.44	0.47
TFIDF1	0.40	0.44	0.46	0.45	0.42
TFIDF2	0.44	0.46	0.45	0.43	0.45

表 4.13: アンカーテキストの重み付けの効果 (k=18 での全属性の平均 F 値の最大値)

(倍率)

	1	5	10	50	100
TF	0.50	0.53	0.57	0.47	0.59
TFIDF1	0.45	0.47	0.51	0.49	0.46
TFIDF2	0.48	0.50	0.54	0.48	0.49

平均 F 値の平均値では全体的に倍率を上げても変動が見られなかった. 局所的に見たところ, 強調倍率 100 で F 値が 0.40〜0.59 へ増加している属性もあったが, 減少している属性もあり, アンカーテキストの重み付けでは大きな分類効果は得られなかった. 平均 F 値の最大値では倍率 10 辺りで F 値が若干高くなっている.

表 4.14, 4.15 に URL の重み付けによる結果を示す.

表 4.14: URL の重み付けの効果 (k=18 での全属性の平均 F 値の平均)

(倍率)

	1	5	10	50	100
TF	0.40	0.46	0.46	0.49	0.51
TFIDF1	0.41	0.41	0.44	0.44	0.47
TFIDF2	0.42	0.44	0.46	0.43	0.47

表 4.15: URL の重み付けの効果 (k=18 での全属性の平均 F 値の最大値)

(倍率)

	1	5	10	50	100
TF	0.43	0.49	0.52	0.57	0.59
TFIDF1	0.44	0.46	0.49	0.48	0.51
TFIDF2	0.43	0.48	0.49	0.48	0.54

ほぼ, 単調増加する結果となった. 局所的に見ると, 減少する属性も存在したが, 0.35〜0.50 へ増加する属性もあり効果的であるといえる.

表 4.16: タイトル、メタタグ、強調タグでの強調の効果 (全て等倍率での F 値の平均)

(倍率)

	1	5	10	50	100
k5	0.34	0.43	0.31	0.38	0.39
k11	0.30	0.44	0.37	0.42	0.44
k18	0.41	0.47	0.43	0.40	0.42

効果は見られなかった. 効果的なタイトルが使用されているサイトが少なかった.(例えば, サイトの文書に対して, タイトルにサイト名しか記述していない) メタタグは対象としたサイトではほとんど用いられていなかった.

次にタイトル, メタタグ, 強調タグのそれぞれで個別に強調操作を行った結果を表 4.17, 4.18, 4.19 に示す. 試行回数はそれぞれ 2 回ずつ行った.

表 4.17: タイトルでの強調効果（全属性の平均 F 値,u1,a1,tf,m1,k1)

(倍率)

	1	5	10	50	100
k5-1	0.32	0.39	0.37	0.43	0.36
k5-2	0.32	0.36	0.39	0.34	0.34
k8-1	0.40	0.39	0.39	0.36	0.41
k8-2	0.44	0.31	0.39	0.43	0.38
k11-1	0.36	0.37	0.47	0.44	0.45
k11-2	0.38	0.42	0.44	0.42	0.45
k15-1	0.44	0.39	0.44	0.40	0.44
k15-2	0.42	0.39	0.43	0.38	0.41
k18-1	0.45	0.44	0.42	0.45	0.50
k18-2	0.42	0.50	0.47	0.42	0.40

※項目説明
k5-1:入力クラスタ数 5, 試行回数 1

クラスタ数を 5, 8, 11, 15, 18 で指定し実験したところ, 試行毎にばらつきがあり, しかも倍率による F 値の相関が明確に表れなかった.

表 4.18: メタタグでの強調効果（全属性の平均 F 値,u1,a1,tf,t1,k1)

(倍率)

	1	5	10	50	100
k5-1	0.32	0.32	0.33	0.32	0.39
k5-2	0.32	0.32	0.33	0.33	0.39
k8-1	0.40	0.33	0.36	0.34	0.34
k8-2	0.44	0.39	0.36	0.37	0.37
k11-1	0.36	0.38	0.37	0.43	0.42
k11-2	0.38	0.37	0.32	0.47	0.36
k15-1	0.44	0.39	0.38	0.39	0.41
k15-2	0.42	0.36	0.40	0.36	0.42
k18-1	0.45	0.40	0.38	0.45	0.39
k18-2	0.42	0.44	0.39	0.42	0.38

※項目説明
k5-1:入力クラスタ数 5, 試行回数 1

メタタグによる強調においてもクラスタ数を 5, 8, 11, 15, 18 で指定し実験したところ, 試行毎に F 値にばらつきがあり, しかも倍率による F 値の相関が明確に表れなかった.

病院 10 サイトでの実験

表 4.19: 強調タグでの強調効果（全属性の平均 F 値,u1,a1,tf,t1,m1)

(倍率)

	1	5	10	50	100
k5-1	0.32	0.33	0.34	0.33	0.30
k5-2	0.32	0.36	0.40	0.33	0.37
k8-1	0.40	0.36	0.42	0.39	0.34
k8-2	0.44	0.38	0.36	0.40	0.35
k11-1	0.36	0.42	0.40	0.40	0.36
k11-2	0.38	0.38	0.42	0.34	0.38
k15-1	0.44	0.34	0.48	0.41	0.39
k15-2	0.42	0.39	0.45	0.47	0.41
k18-1	0.45	0.42	0.38	0.41	0.42
k18-2	0.42	0.41	0.45	0.44	0.44

※項目説明
k5-1:入力クラスタ数 5, 試行回数 1

クラスタ数を 5, 8, 11, 15, 18 で指定し実験したところ, 試行毎にばらつきがあり, しかも倍率による F 値の相関が明確に表れなかった.

表 4.20: 各種パラメータの強調効果 (F 値の平均)

		クラスタ数					ave.
		5	12	20	28	35	
all1	TF	0.15	0.21	0.26	0.31	0.30	0.25
	TFIDF1	0.16	0.20	0.24	0.26	0.28	0.23
	TFIDF2	0.14	0.19	0.24	0.24	0.27	0.22
u100	TF	0.17	0.24	0.31	0.34	0.36	0.28
	TFIDF1	0.16	0.24	0.31	0.34	0.34	0.28
	TFIDF2	0.17	0.25	0.30	0.33	0.33	0.28
a100	TF	0.17	0.24	0.30	0.31	0.35	0.27
	TFIDF1	0.18	0.24	0.30	0.35	0.35	0.28
	TFIDF2	0.17	0.25	0.31	0.30	0.34	0.27
t100	TF	0.16	0.26	0.26	0.27	0.29	0.25
	TFIDF1	0.16	0.24	0.23	0.29	0.29	0.24
	TFIDF2	0.14	0.20	0.25	0.28	0.27	0.23
m100	TF	0.17	0.23	0.29	0.29	0.31	0.26
	TFIDF1	0.16	0.19	0.24	0.25	0.27	0.22
	TFIDF2	0.16	0.19	0.23	0.25	0.25	0.22
k100	TF	0.18	0.24	0.28	0.31	0.34	0.27
	TFIDF1	0.17	0.22	0.26	0.28	0.35	0.26
	TFIDF2	0.15	0.21	0.26	0.26	0.32	0.24

all1:全てのパラメータで倍率を 1
u100:URL の強調倍率を 100
a:アンカーテキストの強調倍率を 100
t100:タイトルの強調倍率を 100
m100:メタ情報の強調倍率を 100
k100:強調文字の倍率を 100

all1(全てのパラメータで倍率を 1 に設定) での条件と比較して見ると, u100(アンカーテキストでの強調) では TF が 0.03, TFIDF1 では 0.05, TFIDF2 では 0.06 増加している. 同様に a100, k100 も増加しているが, t100, m100 では TF, TFIDF1, TFIDF2 によっては減少しているものも見られた.

ここで予備実験で行った属性候補の自動抽出の属性名に沿い抽出した異表記同義語関係を以下に示す. 予備実験で行ったクラスタ数 12, F 値が最も高かったパラメータの条件としたクラスタリング結果から抽出した異表記同義語関係を以下に示す.

表 4.21: 抽出された異表記同義語の例

クラスタ	異表記同義語
1	バランス生活 第 10 号 第 9 号 第 7 号 広報誌ありまつ第 6 号を追加しました (2005/8/20) 広報誌ありまつ第 5 号を追加しました (2005/6/3) 第 11 号 2007/2/3 更新 案内地図
2	肩関節専門外来のページへ 肩関節専門外来を開設しております 肩関節専門外来 【内視鏡センターのページへ戻る】 胃腸科 内視鏡センター
3	交通アクセス このページの先頭へ アクセス お問合せ このページの先頭へ 作業療法と病院行事 施設のご案内 診療科のご案内 診療科 サイトマップ 小松市民病院 (head) 管理局 各種教室 地図、交通手段 個人情報の利用目的 ページトップへ トップページ 小松市民病院 (body) バス時刻表 案内地図 戻る ディケアセンター 栄養一口メモ RETURN 内科 RETURN 循環器科

表 4.22: 抽出された異表記同義語の例

クラスタ	異表記同義語
	あなたは ※各種クレジットが使えます。詳しくは、受付の方へお尋ね下さい。 お支払はカードがご利用できます こちら 健康教室 各種教室案内 地域連携室 地域医療連携室 職員募集情報 サイトマップ 地図・アクセス方法 DoCoMo 戻る 公立松任石川中央病院 (fraTop) J-PHONE 戻る 各種診断書 ＰＥＴセンター アクセス 院内行事 HOME 公立つるぎ病院 (main) 無題ドキュメント (tab) 項目 料金
クラスタ	異表記同義語
4	診療科 HP リンク集 このページの先頭へ 診療のご案内 このページの先頭へ 診療科一覧 診療のご案内 電話番号一覧 このページの先頭へ 診療科案内　診療科一覧 診療科 診療科案内 受診手続き このページの先頭へ 外来診察（初診・再診）の手続き方法 ご挨拶 診療科目、診療時間 バス時刻表

表 4.23: 抽出された異表記同義語の例

クラスタ	異表記同義語
	小松市民病院 (menu) 診療所 なかいクリニック 戻る RETURN 外来診療時間変更のお知らせ 外来診療時間のお知らせ 外来診療時間 2007/2/9 更新 診療時間 診療科・受付時間 戻る インフォメーション
5	看護師募集 ページの先頭に戻る 平成 19 年度 看護師募集要項 平成 19 年度看護師募集 こちらからお入り下さい。 訪問看護ステーションのページへ 求人情報 戻る 介護支援事業 訪問看護ステーション 看護師募集
6	医療関係の方 お知らせとお願い このページの先頭へ ８月１５日（火） 外来診療休診のお知らせ このページの先頭へ グループホーム グループホーム「ハイツ北金沢」 詳細ページへ 福祉ホームＢ型 詳細ページへ 循環器科 泌尿器科 院長の挨拶 – 医療法人 社団 和楽仁　芳珠記念病院 –() お知らせ 戻る 広報誌ヤッ芳

表 4.24: 抽出された異表記同義語の例

クラスタ	異表記同義語
	What's New!
	ＨＯＭＥへ
	金沢有松病院ホームページ
	皮膚科
	人間ドックの案内
	健康トピックスに「乳がんが増えています」
	人間ドック
	麻酔科
	ケミカルピーリング (MicroPeel)
	ケミカルピーリングのご案内
	全館禁煙のご協力とお願い
	NST とは？
	NST 活動をしています (2005/5/30)
	NST 活動をしています
	フォトフェイシャルのご案内
	フォトフェイシャル (Photofacial)
	RETURN
	最新機器の紹介
	整形外科
	日本皮膚科学会認定専門医研修施設
	日本静脈経腸栄養学会栄養サポートチーム専門療法士認定規則実地修練認定教育施設
	学会認定施設など一覧
	禁煙外来のお知らせ
	泌尿器科
	広報誌（医療 NOW）
	広報誌&医療 NOW
	お見舞いメール
	過去のお知らせ
	更新履歴一覧
	More
	理念
	院内施設
	ＰＥＴ検査を開始しました
	ピンクリボンキャンペーン実施中
	公立松任石川中央病院 (fraMain)
	各部署紹介
	縁の下の力持ち
	公立つるぎ病院 (head)
	外来診察担当一覧表
	外来診察表

表 4.25: 抽出された異表記同義語の例

クラスタ	異表記同義語
	外来担当医一覧表 外来診察一覧 無題ドキュメント (home_main) 病院長所感 リンク集
7	花便り 消化器科 外科 H18 年 10 月より第 1・3・5 木曜日午後に呼吸器外科外来を開設しております 医師紹介 外来案内 概要と沿革 ページトップへ 概要 医療技術部から 各診療科 曜日別各科外来医師表 人間ドック（健診） 初めて来院の方 外来診療の案内 病院概要 標榜科 New 患者様の権利に関する宣言 プライバシーポリシー ご意見＆お問い合せ ご意見＆お問い合わせ 医療のはなし 受付方法 外来には以前に行ったことある 紹介状をお持ちの方 はじめての方 はじめて外来に行く。 健康診断 概要 概要・特色 連携医療機関 連携医療機関名、医師名一覧 理解が深まる病院情報 職員募集 病院の概要 ご意見箱 基本理念 スタッフ紹介

表 4.26: 抽出された異表記同義語の例

クラスタ	異表記同義語
8	施設案内
	病院長ご挨拶
	病院長
	病院の概要
	本院の基本理念
	平成 19 年度金沢大学病院歯科研修プログラムのお知らせ
	平成 19 年度金沢大学病院歯科研修プログラム
	金沢大学病院モニター募集のお知らせ
	入院案内
	入院のご案内
	←目次へ
	地域医療連携インフォメーション
	専門医一覧
	診療科案内　専門医一覧
	卒後臨床研修センター
	病院へのアクセス
	モーツアルトの音楽のような医療を目指して
	当院の精神科医療への考え方
	病院のご案内
	受診と入院のご案内
	平成 19 年度 北陸病院後期臨床研修募集要項
	入院案内
	入院案内
	芳珠記念病院
	外科
	肛門科
	呼吸器科
	金沢有松病院ホームページ目次
	Kanazawa Arimatsu hospital(window1)
	脳神経外科
	病院概要
	公立松任石川中央病院 (fraMenu)
	入院の前に
	お見舞い
	病院長あいさつ
	病院長顔写真
	メニュー
	総合メニュー
	平成１９年度初期臨床研修医の募集を行います。
	●平成 19 年採用後期研修医採用試験案内
	募集要項
	公立つるぎ病院 (contents)

表 4.27: 抽出された異表記同義語の例

クラスタ	異表記同義語
	健診・ドック
	健診案内
	一般健診・ドックのご案内
	検診案内
	病院の特徴
	受付
9	金沢大学医学部附属病院
	臨床試験監理センター
	DPC 関連資料
	臨床試験管理センター
	専門医養成コースについてのお知らせ
	専門医養成コース
	放射線科
	個人情報保護方針
	メディカルエステ ARIMATSU
	メディカルエステのご案内
	メディカルエステ ARIMATSU のご案内
	個人情報取扱いについて (2005/3/31)
	プライバシーポリシー
	●採用試験案内 (1 月採用)
	●採用試験案内 (4 月採用)
	患者様の個人情報保護について
	病院における個人情報の保護について
10	⇒看護部ホームページ
	金沢大学医学部附属病院 看護部
	看護部
	看護部から
	看護部教育指針
	院内報てどり
11	個人情報保護
	患者さまへ
	患者様へ
	地域医療連携室
	↓当院での診療を希望される患者様へ
	↓当院に入院及び通院中の患者様へ
	↓地域医療機関の皆様へ
	デイケア＆デイナイトケア
	デイケア＆デイナイトケア「さくらんぼ」
	詳細ページへ
	福祉ホームＢ型「プリムラ」
	個人情報の取り扱いについて
	当院における個人情報保護に関する方針
	地域医療連携室から

表 4.28: 抽出された異表記同義語の例

クラスタ	異表記同義語
	健診センター (人間ドック) 『健診内容をご覧ください。』 こちらからご覧ください 健診センター 基本方針 施設紹介 お薬の豆知識 RETURN 地域医療連携室開設 財団法人日本医療機能評価機構　認定病院となりました 病院機能評価認定について 透析センター 入院案内 総合健診センター 本文へジャンプ 総合検診センター 検診オプション アクセスマップ 人間ドック 高額療養費制度 各種公的制度 各種制度 石川県乳幼児助成制度 乳幼児助成金制度 乳幼児助成 更新履歴
12	レーザー脱毛のご案内 レーザ脱毛のご案内 医療レーザー脱毛とは LightSheer とは 医療用ダイオードレーザー「LightSheer」

それぞれのクラスタに着目すると, 予備実験で得られた属性項目でのクラスタリング結果とならず, 「肩関節」に関するクラスタ, 「広報」に関するクラスタ, 「看護」に関するクラスタに分別される結果となった.

4.5 考察

本研究ではクラスに対する対象物から属性情報を抽出する手法を提案した. 具体的には Web 文書を分類し, アンカーテキストの異表記同義語関係の抽出を試みた.

アンカーテキストにおける異表記同義語関係の抽出手法として, 本研究ではベクトル空間法を用い非階層型のクラスタリング手法 k-means で文書を分類することで, それら文書を示すアンカーテキストを収集し実現しようとした. 索引語の頻度に対する重み付けを URL, アンカーテキスト, タイトル, メタタグ, 強調文字に対して行うことである程度の効果を示すことができた. その各種パラメータの倍率を上げクラスタリングを行うことで, 平均 F 値を 0.01〜0.05 の向上させることができたが, 捉え方によっては微々たるものに過ぎないことは否定できない.

原因として, まずクラスタリングの初期割り当てでランダムなセントロイドを設定しサンプルデータをクラスタに割り当てていることが考えられる. ランダムに行うことの影響は強く, 属性毎の分類 (意味分類) ではなく, サイト特有の索引語の影響を強く受け, 特定のサイトの文書で集まっているクラスタが見られた. 如何に意味分類を追求するため, WordNet による類義語表現, 英和辞書を活用するなど行ったが, 大きな効果を得るには至っていない.

次に, 各種パラメータを同時に強め合うことでの競合が考えられる. 実験データから, 全てのパラメータを最適な F 値から求めた倍率に設定したとしても, 単純に一種のパラメータのみの場合よりも F 値が低くなる場合があった. 比較的 URL による強調が他の強調より効果的であったため, URL 強調の効果を活かせるパラメータ設定が重要である.

また, $TF \cdot IDF$ による重み付けを行ったが, 本実験では効果を得ることができなかった. 原因として, 属性語が一般的な表現であることが多いため, 文書の特徴となる索引語の重みを強くする $TF \cdot IDF$ 値では, 同サイトの文書同士がクラスタリングで集まりやすくなったことが考えられる. (例えば, 特徴的な単語 “北陸先端大” が強まり, 北陸先端大に属する研究室が同じクラスタに集まりやすくなる. 研究室クラスでは研究やメンバーなどに関する文書が集まるクラスタが生成されることが好ましいが, $TF \cdot IDF$ により “研究” や “メンバー” という単語は強調されうる単語とは推測され難い.)

第5章　おわりに

5.1　まとめ

本稿では, 与えられたクラスの対象物を記述した Web サイト (Web ページの集合) のいくつかの対象物から, HTML タグを解析し, 異表記同義語関係の抽出方法を提案し, 実験によりその有効性を示した.

具体的には, Web 文書からアンカーテキスト, URL, タイトル, メタ情報, 強調文字を利用し, 本文から生成した索引語ベクトルの重み付けを行うことで, 異表記同義語関係を抽出を行った. 実験によって提案手法を用い, 対象物の異表記同義語関係を抽出可能であることを示した.

5.2　今後の課題

今後の課題を以下に述べる.

まず, クラスに属する属性の特定が必要である. 今回は, Web 文書からの索引語ベクトルの生成と基本的なクラスタリング手法 k-means のアルゴリズムをそのまま実装したに過ぎず, 初期のクラスタへの割り当てで工夫の余地が考えられた. Step3 で抽出した, URL からの属性候補を使用し, 初期のクラスタ数やサンプルデータの初期クラスタへの割り当ててを行ってからクラスタリングさせることで精度を向上させることが可能と考える.

今回実装しやすさと高速性から非階層型のクラスタリング手法を採用したが, 入力データの規模など場合によっては階層型のクラスタリングでも効果が見られると思われる. また, 今回着目した各種パラメータのみで単純なマッチングによりクラスタリングさせその効果を確認し, その後, 索引語ベクトルを生成し再度クラスタリングされる手法も考えられる

次に複数人による属性決定と評価を行う必要がある. 今回の実験では, 時間の都合により, 実験者の主観によりクラスの属性 (例. 入院案内, 研究) を同定した. しかし, 初期値に結果が大きく左右される本クラスタリング手法では複数人で同定した属性もしくはそれに相当する属性を入力することでクラスタリングを行うことにより精度が向上できることが予想される.

最後に, クラスタリングの高速化の検討が必要である. クラスにより大きなばらつきがあるが, 実験結果で示したサンプルデータを入力値をした場合, 病院 5 サイトのクラスタリング処理に時間が費やされた. 研究室のサイトと比較し本文の文字数が多いことで, 抽

出される索引語候補数が増えることが原因と考えられるが, 文字を中心に記述される傾向の強いクラスほど計算コストがかかる. この点については徳永 [12][1]にも記載されているように “出現頻度の上限と下限を設け, その敷値を超える高頻度語と低頻度語を索引語として採用しないのが一般的” である. 索引語を絞ることで必要な記憶領域や計算速度の面でメリットがあるが, 本研究で対象とした Web 文書に対しては一概に絞ることは難しいと思われた. 例えば, 研究に関する文書では “研究” や “research” など属性語に関わる索引語候補の出現頻度は高くなる傾向にあったが, 交通案内に関するページでは, “交通” や “access” など属性語に関わりの強い語に対し出現頻度が 0 や 1 で, 敷値を定めるのは困難である. よって, 高速化を図るには URL やアンカーテキストなど, 今回対象とした各種パラメータから索引語候補を求めることで, 本文から生成した索引語数に比べ少なくなり, 小さな次元でのベクトル空間を対象とすることができる.

[1] p.22

謝辞

本研究室を進めるにあたり鳥澤健太郎助教授には, 日頃より研究方針や研究内容など研究全般につきまして, 有益なご意見・ご助言を頂くなど非常に熱心なご指導を賜りました.

そして, 基礎知識から実装まで幅広く様々な助言を頂いた, 吉永直樹氏にはこの場を借りて暑く御礼申し上げます.

また, 本研究にご理解と多大なご協力を賜りました, 東条敏教授をはじめとする知識工学講座の皆様方に深く感謝するとともに, 特に食の面からサポートして頂いた二本真氏, 研究活動から生命活動まで力添えを頂いた松永博充氏, マンネリとしていた研究活動に彩りを与えてくれた高崎晃一氏には心から感謝いたします.

参考文献

[1] 赤穂 昭太郎, 神嶌 敏弘ら. 朱鷺の杜. http://www.neurosci.aist.go.jp/ibisforest/.

[2] 鳥澤 健太郎 徳永 耕亮, 風間 淳一. 上位下位関係を用いた html 文書からの属性及び属性値の自動抽出. **言語処理学会 第 *11* 回年次大会**, 2005.

[3] 獅々堀 正幹 北　研二, 津田 和彦. **情報検索アルゴリズム**. 共立出版株式会社, 2002.

[4] 中川 裕志 湯本 紘彰, 森辰則. 出現頻度と連接頻度に基づく専門用語抽出. **自然言語処理**, 10(1):27–45, 2003.

[5] 鈴木祐介, 松原茂樹, and 吉川正俊. アンカーテキストとハイパーリンクに基づく Web 文書の階層的分類. **人工知能学会第 *19* 回全国大会論文集**; *3C2-02*, 2005.

[6] 鈴木祐介, 松原茂樹, and 吉川正俊. アンカーテキストを用いた Web ディレクトリの構築. **電子情報通信学会技術研究報告**, 105(203), 2005.

[7] 小松 弘幸. ローマ字ひらがな変換ライブラリ suikyo. http://taiyaki.org/.

[8] 工藤 拓. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.jp/.

[9] まつもとゆきひろ. オブジェクト指向スクリプト言語 Ruby. http://www.ruby-lang.org/.

[10] 松本裕治 高橋哲朗, 乾健太郎. テキストから属性情報を抽出する. **情報処理学会研究報告**, **自然言語処理研究会**, pages 19–24, 11 2004.

[11] 高林哲. Migemo. http://0xcc.net/migemo/.

[12] 徳永 健伸. **情報検索と言語処理** *(***言語と計算** *5)*. 東京大学出版会, 1999.

[13] 原 信一郎. *Ruby* **プログラミング入門**. 株式会社 オーム社, 2000.

[14] 鳥澤健太郎 新里圭司. Html 文書からの単語間の上位下位関係の自動獲得. **自然言語処理**, 12:–, 2003.

[15] 新里圭司. 鳥澤健太郎. Html 文書中の箇条書きとその表題に注目した下位語の自動獲得. **情報処理学会**, **研究報告**, pages –, 2004.

[16] 佐々木稔 新納浩幸. Web ページ内の目的部分の自動抽出. **情報処理学会**, **研究報告**, pages –, 2004.

[17] 鳥澤健太郎 吉永直樹. Web からの属性情報記述ページの発見. **言語処理学会第 *12* 回年次大会**, pages –, 2006.

[18] Tom Mitchell Avrim Blum. Combining labeled and unlabeled data with co-training. *Proceedings of the 1998 Conference on Computational Learning*, 98.

[19] Free Software Foundation, Inc. GNU Wget - GNU Project - Free Software Foundation (FSF). http://www.gnu.org/software/wget/wget.html.

[20] George A. Miller. WordNet - Princeton University Cognitive Science Laboratory. http://wordnet.princeton.edu/.

[21] Hiroshi Nakagawa, Akira Maeda and Hiroyuki Kojima. 専門用語 (キーワード) 自動抽出システム. http://gensen.dl.itc.u-tokyo.ac.jp/.

[22] Incept Inc. HTML の特殊文字 (IT 用語辞典 e-Words). http://e-words.jp/p/r-htmlentity.html.

[23] J.B.MacQueen. Some methods of classification and analysis of multivariate observations. *Proc. of 5th Berleley Symposium on Math. Stat. and Prob.,pp.281-297*, 1967.

[24] M.Y. Kan and H.O.N. Thi. Fast webpage classification using URL features. *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 325–326, 2005.

[25] Kurumi. Gene95 辞書. http://www.namazu.org/%7Etsuchiya/sdic/data/gene.html.

[26] M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically synthesizing web pages. *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 727–732, 1998.

[27] RubyForge. RubyForge. http://rubyforge.org/.

[28] K. Shinzato and K. Torisawa. Acquiring hyponymy relations from web documents. *Proceedings of HLT-NAACL*, 80, 2004.

[29] SKK Openlab. SKK Openlab. http://openlab.jp/skk/index-j.html.

[30] Sussex NLCL. Sussex NLCL. http://www.informatics.susx.ac.uk/research/nlp/.

[31] W3C, HTML4 仕様邦訳計画補完委員会. W3C 勧告私的日本語翻訳. http://www.asahi-net.or.jp/%7Esd5a-ucd/rec-html401j/cover.html.

[32] Yahoo! Japan. Yahoo! デベロッパーネットワーク. http://developer.yahoo.co.jp/.

[33] ZSPC. ウェブコンテンツ・アクセシビリティ・ガイドライン 1.0. http://www.zspc.com/documents/wcag10/.

付録

表 5.1: 条件 1 の正解データ

属性	URL
member	www-oml.sum.sd.keio.ac.jp/jp/members/index.html
	kt-www.jaist.ac.jp/kt-intro/members-j.html
	www.tqm.t.u-tokyo.ac.jp/member/member.html
	kt-www.jaist.ac.jp/kt-intro/ob/ob-2000.html
	kt-www.jaist.ac.jp/kt-intro/ob/ob-2001.html
	kt-www.jaist.ac.jp/kt-intro/ob/ob-2002.html
	kt-www.jaist.ac.jp/kt-intro/ob/ob-2003.html
	kt-www.jaist.ac.jp/kt-intro/members.html
	www-tori.jaist.ac.jp:8000/member.html
	www-tori.jaist.ac.jp:8000/member.en.html
	www-oml.sum.sd.keio.ac.jp/en/members/index.html
	kt-www.jaist.ac.jp/kt-intro/ob/ob-2004.html
	www.r.dl.itc.u-tokyo.ac.jp/members-e.html
kojin	kt-www.jaist.ac.jp/ toshiaki/index-j.html
	www.r.dl.itc.u-tokyo.ac.jp/ nakagawa/
	www.r.dl.itc.u-tokyo.ac.jp/ nakagawa/index.html
	www.r.dl.itc.u-tokyo.ac.jp/ yositomi/
	www.r.dl.itc.u-tokyo.ac.jp/ mino/index.htm
	www.r.dl.itc.u-tokyo.ac.jp/ nakagawa/English.html
	www.r.dl.itc.u-tokyo.ac.jp/ kiyota/index-ja.html
	www.r.dl.itc.u-tokyo.ac.jp/ haraguchi/index.html
	www.r.dl.itc.u-tokyo.ac.jp/ hoshino/
	www.r.dl.itc.u-tokyo.ac.jp/ kiyota/index-en.html

属性	URL
project	www-tori.jaist.ac.jp:8000/project/
	www-tori.jaist.ac.jp:8000/project/index.html
	kt-www.jaist.ac.jp/project/index.html
access	www-oml.sum.sd.keio.ac.jp/jp/about/index.html
	www.tqm.t.u-tokyo.ac.jp/access/access.html
	kt-www.jaist.ac.jp/kt-intro/map.html
	kt-www.jaist.ac.jp/kt-intro/address-j.html
	kt-www.jaist.ac.jp/kt-intro/address-e.html
	www.r.dl.itc.u-tokyo.ac.jp/access-e.html
	kt-www.jaist.ac.jp/kt-intro/map-e.html
research	www-oml.sum.sd.keio.ac.jp/jp/research/index.html
	www-tori.jaist.ac.jp:8000/theme.html
	kt-www.jaist.ac.jp/research/index-j.html
	www.tqm.t.u-tokyo.ac.jp/research/research.html
	www-tori.jaist.ac.jp:8000/publications.en.html
	www.r.dl.itc.u-tokyo.ac.jp/research_topics-e.html
	www-oml.sum.sd.keio.ac.jp/jp/research/list/2005.html
	www-tori.jaist.ac.jp:8000/publications.html
	www-tori.jaist.ac.jp:8000/theme.en.html
	kt-www.jaist.ac.jp/research/index.html
	www-oml.sum.sd.keio.ac.jp/en/research/index.html
	www-oml.sum.sd.keio.ac.jp/jp/research/bilateral/index.html
	www-oml.sum.sd.keio.ac.jp/jp/research/humanoid/index.html
	www.r.dl.itc.u-tokyo.ac.jp/works-e.html
	www-tori.jaist.ac.jp:8000/papers/jsps99-hpsg.pdf
top	www-tori.jaist.ac.jp:8000/index.html
	www-oml.sum.sd.keio.ac.jp/jp/index.html
	www.r.dl.itc.u-tokyo.ac.jp/index.html
	www.tqm.t.u-tokyo.ac.jp/index.html
	kt-www.jaist.ac.jp/index-j.html
	www-oml.sum.sd.keio.ac.jp/en/index.html
	www-oml.sum.sd.keio.ac.jp/jp/about/ohnishi.html
	kt-www.jaist.ac.jp/index-e.html
	www.r.dl.itc.u-tokyo.ac.jp/index-e.html

属性	URL
other	www-tori.jaist.ac.jp:8000/meeting/
	www-tori.jaist.ac.jp:8000/meeting/index.html
	www-tori.jaist.ac.jp:8000/topics.html
	www-tori.jaist.ac.jp:8000/dsearch/
	www-oml.sum.sd.keio.ac.jp/en/download/index.html
	www-oml.sum.sd.keio.ac.jp/jp/download/index.html
	www-oml.sum.sd.keio.ac.jp/en/about/admissions.html
	kt-www.jaist.ac.jp/cb4office/office.cgi
	www-oml.sum.sd.keio.ac.jp/jp/about/admission.html
	www-tori.jaist.ac.jp:8000/cgi-bin/bbs/bbs.cgi
	www.tqm.t.u-tokyo.ac.jp/link/link.html

表 5.2: 条件 2 の正解データ

属性	URL
toiawase	web.hosp.kanazawa-u.ac.jp/others/otoiawase.html
	www.kanazawa-rc-hosp.jp/toiawase/index.html
link	web.hosp.kanazawa-u.ac.jp/others/kouzalink.html
	www.mattohp.jp/hi/i_openhp.html
	www.tsurugihp.jp/chikaramochi/index.html
	www.tsurugihp.jp/link/index.html
access	web.hosp.kanazawa-u.ac.jp/others/kotsu.html
	www.e-sakurahp.com/access/index.html
	www.hosp.komatsu.ishikawa.jp/map_access.html
	www.houju.or.jp/bus.html
	www.houju.or.jp/map.html
	www.knh.or.jp/newpage4.htm
	www.mattohp.jp/an/a_access.html
	www.tsurugihp.jp/access/index.html
privacy	web.hosp.kanazawa-u.ac.jp/others/kojin_jyoho.html
	www.hokuriku-hosp.jp/personal_data.htm
	www.hosp.komatsu.ishikawa.jp/privacy.html
	www.hosp.komatsu.ishikawa.jp/riyo_mokuteki.html
	www.kanazawa-rc-hosp.jp/puraibasi/index.html
	www.k-arimatsu.jp/kenri.html
	www.k-arimatsu.jp/privacy-houshin.html
	www.mattohp.jp/qa/q_ho0006.html
bosyu,saiyou	web.hosp.kanazawa-u.ac.jp/bu/kango/05bosyu/bosyu.html
	www.hokuriku-hosp.jp/info/kensyuui_bosyu.htm
	www.hosp.komatsu.ishikawa.jp/kanrikyoku.html
	www.houju.or.jp/joblist.htm
	www.kanazawa-rc-hosp.jp/saiyou/index.html
	www.knh.or.jp/bosyuu.htm
	www.mattohp.jp/pc/k_bosyu_06.html
	www.mattohp.jp/pc2/k2_bosyu_06.html
	www.mattohp.jp/pu/bosyu_06-2.html
	www.mattohp.jp/pu/bosyu_06-N.html
	www.tsurugihp.jp/bosyuu/index.html

属性	URL
oshirase, rireki	web.hosp.kanazawa-u.ac.jp/gallery/gallery36.html
	web.hosp.kanazawa-u.ac.jp/oshirase/050920.html
	web.hosp.kanazawa-u.ac.jp/oshirase/H19shika.html
	web.hosp.kanazawa-u.ac.jp/oshirase/hm070201.html
	web.hosp.kanazawa-u.ac.jp/oshirase/index.html
	web.hosp.kanazawa-u.ac.jp/oshirase/kyushin060815.html
	www.houju.or.jp/news.html
	www.kanazawa-rc-hosp.jp/infomation/20060308info.html
	www.kanazawa-rc-hosp.jp/infomation/20060819info.html
	www.kanazawa-rc-hosp.jp/infomation/20060929info.html
	www.kanazawa-rc-hosp.jp/infomation/20061201info.html
	www.kanazawa-rc-hosp.jp/rireki/index.html
	www.k-arimatsu.jp/info_kyuusin.html
	www.k-arimatsu.jp/kenko.html
	www.k-arimatsu.jp/smoke.html
	www.mattohp.jp/rireki.html
event	www.e-sakurahp.com/gyouji/index.html
	www.tsurugihp.jp/gyouji/index.html
sisetsu	web.hosp.kanazawa-u.ac.jp/annai/index.html
	web.hosp.kanazawa-u.ac.jp/bu/yaku/crc/
	web.hosp.kanazawa-u.ac.jp/sotsuken/
	www.e-sakurahp.com/haitsu/index.html
	www.e-sakurahp.com/purimura/index.html
	www.e-sakurahp.com/sakuranbo/index.html
	www.e-sakurahp.com/shisetsu/index.html
	www.hosp.komatsu.ishikawa.jp/kakuka/kensin/index.html
	www.houju.or.jp/sisetu.html
	www.k-arimatsu.jp/announce.html
	www.k-arimatsu.jp/naisikyo.html
	www.k-arimatsu.jp/nst-1.html
	www.k-arimatsu.jp/shisetsu/shisetsu.html
	www.k-arimatsu.jp/touseki/touseki.html
	www.mattohp.jp/hi/i_sisetu.html
	www.mattohp.jp/kenshin/index.html
	www.mattohp.jp/pet/index_c.html
	www.mattohp.jp/pet/index.html

属性	URL
gaiyou, aisatsu	web.hosp.kanazawa-u.ac.jp/gaiyou/incho.html
	web.hosp.kanazawa-u.ac.jp/gaiyou/index.html
	web.hosp.kanazawa-u.ac.jp/gaiyou/rinen.html
	www.e-sakurahp.com/bgm/index.html
	www.e-sakurahp.com/aisatsu/index.html
	www.e-sakurahp.com/byouin/index.html
	www.hosp.komatsu.ishikawa.jp/gaiyo.html
	www.hosp.komatsu.ishikawa.jp/incho.html
	www.houju.or.jp/gaiyou.html
	www.houju.or.jp/plan.html
	www.knh.or.jp/gaiyo.html
	www.mattohp.jp/hi/i_boss.html
	www.mattohp.jp/hi/i_gaiyou.html
	www.mattohp.jp/hi/i_rinen.html
	www.tsurugihp.jp/gaiyou/index.html
	www.tsurugihp.jp/home/infomation.html
	www.tsurugihp.jp/intyou/index.html
	www.tsurugihp.jp/rinen/index.html
	www.tsurugihp.jp/tokutyou/index.html
kouhou	web.hosp.kanazawa-u.ac.jp/diet/index.html
	www.houju.or.jp/pr.html
	www.kanazawa-rc-hosp.jp/kohoshi/index.html
	www.k-arimatsu.jp/kouhou_arimatsu/kouhou.html
	www.tsurugihp.jp/tedori/index.html
nyuin	web.hosp.kanazawa-u.ac.jp/patients/nyuin/index.html
	www.hosp.komatsu.ishikawa.jp/nyuin.html
	www.houju.or.jp/nyuin.html
	www.knh.or.jp/nyuin1.htm
	www.mattohp.jp/gn/n_info.html
	www.mattohp.jp/gn/n_omimai.html

属性	URL
shinryou, jusin, gairai	web.hosp.kanazawa-u.ac.jp/patients/guide/index.html web.hosp.kanazawa-u.ac.jp/patients/index.html web.hosp.kanazawa-u.ac.jp/patients/info/list.html web.hosp.kanazawa-u.ac.jp/patients/tetsuzuki/index.html www.e-sakurahp.com/jyushin/index.html www.hokuriku-hosp.jp/visitor/shoulder_joint.htm www.hosp.komatsu.ishikawa.jp/gairai.html www.houju.or.jp/first.html www.houju.or.jp/gairai.html www.knh.or.jp/sinryou2.htm www.mattohp.jp/an/s_sinryo.html www.mattohp.jp/gn/g_saisin.html www.mattohp.jp/gn/g_syokai.html www.mattohp.jp/gn/g_syosin.html www.mattohp.jp/gn/k_menu.html
renkei, kyousitu	web.hosp.kanazawa-u.ac.jp/patients/renkei/index.html web.hosp.kanazawa-u.ac.jp/patients/renkei/relpert.html www.hosp.komatsu.ishikawa.jp/fukusisitu.html www.hosp.komatsu.ishikawa.jp/kyositu.html www.hosp.komatsu.ishikawa.jp/sinryojo.html www.houju.or.jp/nakai/nakai.html www.kanazawa-rc-hosp.jp/kenkou/index.html www.kanazawa-rc-hosp.jp/renkei/index.html www.k-arimatsu.jp/eria-iryo.html
member	web.hosp.kanazawa-u.ac.jp/senmoni/index.html www.hosp.komatsu.ishikawa.jp/doctor.html www.hosp.komatsu.ishikawa.jp/weekly.html www.tsurugihp.jp/staff/index.html
sitemap,menu	www.hokuriku-hosp.jp/site_map.htm www.kanazawa-rc-hosp.jp/sitmap/index.html www.k-arimatsu.jp/mokuji.html www.mattohp.jp/menu.html www.mattohp.jp/f_menu.html www.hosp.komatsu.ishikawa.jp/sidemenu.html www.houju.or.jp/menu.html www.tsurugihp.jp/contents.html www.tsurugihp.jp/uketuke/index.html

属性	URL
sinryouka	www.hokuriku-hosp.jp/medical_examination/circulatory.htm
	www.hokuriku-hosp.jp/medical_examination/digestive.htm
	www.hokuriku-hosp.jp/medical_examination/medical_examination.htm
	www.hokuriku-hosp.jp/medical_examination/radiation.htm
	www.hokuriku-hosp.jp/medical_examination/surgery.htm
	www.hokuriku-hosp.jp/medical_examination/urinary.htm
	www.hosp.komatsu.ishikawa.jp/kakuka/index.html
	www.hosp.komatsu.ishikawa.jp/kamoku_jikan.html
	www.k-arimatsu.jp/datumou.html
	www.k-arimatsu.jp/dermatolo.html
	www.k-arimatsu.jp/drug.html
	www.k-arimatsu.jp/eiyou.html
	www.k-arimatsu.jp/estetic/medical.html
	www.k-arimatsu.jp/geka.html
	www.k-arimatsu.jp/hyobouka.html
	www.k-arimatsu.jp/info.html
	www.k-arimatsu.jp/naika.html
	www.k-arimatsu.jp/nouge.html
	www.k-arimatsu.jp/seikei.html
	www.tsurugihp.jp/busyo/index.html
	www.tsurugihp.jp/home/gairai.html
houmon	www.hokuriku-hosp.jp/station/station.htm
	www.kanazawa-rc-hosp.jp/kaigo/houmon/index.html
top	www.hosp.komatsu.ishikawa.jp/toppage.html
	www.houju.or.jp/top.html
	www.k-arimatsu.jp/arimatsu.html
	www.mattohp.jp/docomo/index.html
	www.mattohp.jp/j-sky/index.html
	www.mattohp.jp/top.html
	www.tsurugihp.jp/home_index.html
	www.tsurugihp.jp/home_main.html
doc	www.houju.or.jp/dock.html
	www.k-arimatsu.jp/dock/dock.html
	www.k-arimatsu.jp/kensin.html
	www.tsurugihp.jp/kensin/index.html

属性	URL
other	web.hosp.kanazawa-u.ac.jp/bu/kango/index.html
	www.hosp.komatsu.ishikawa.jp/gijutubu/index.html
	www.hosp.komatsu.ishikawa.jp/head.html
	www.hosp.komatsu.ishikawa.jp/kangobu.html
	www.kanazawa-rc-hosp.jp/omimai_mail/index.html
	www.k-arimatsu.jp/micropeel/micropeel-1.html
	www.k-arimatsu.jp/nintei/nintei.html
	www.k-arimatsu.jp/nosomke.html
	www.k-arimatsu.jp/office_debit.html
	www.k-arimatsu.jp/photofacial/photofacial-1.html
	www.k-arimatsu.jp/saisinkiki.html
	www.k-arimatsu.jp/urologic.html
	www.mattohp.jp/co/index.html
	www.mattohp.jp/f_hpinfo.html
	www.mattohp.jp/op/b_sindan.html
	www.mattohp.jp/op/s_kougak.html
	www.mattohp.jp/op/s_menu.html
	www.mattohp.jp/op/s_nyuyou.html
	www.mattohp.jp/pu/p_ribbon05.html
	www.mattohp.jp/qa/q_hosoku.html
	www.tsurugihp.jp/goiken/index.html
	www.tsurugihp.jp/head.html
	www.tsurugihp.jp/home_tab.html
	www.tsurugihp.jp/ryoukin/index.html

表 5.3: 自動抽出された属性名 (自治体 10 サイト)

about	access	acrobat	action	actionplan	address	air
annai	anzen	asbesto	autumn	backnumber	bosyuu	bousai
brand	bunka	bunkazai	bus	business	byouin	calendar
camp	campaign	center	check	chiji	chinese	chiseki
circle	city	conference	content	culture	daigaku	data
dayori	default	demae	dentou	design	detail	doboku
doctor	douken	douro	download	eco	eisei	eizen
ekisyuhen	english	enquete	entry	etc	event	faq
font	forum	fukushi	funin	fureai	furusato	gairai
gaiyou	gakkou	gallery	gappei	gesui	get	gikai
goiken	gomi	green	guidance	guide	gyakutai	gyoji
gyoukaku	gyoumu	gyousei	hajimeni	hakusanroku	hall	handbook
header	health	help	henkou	hiroba	history	hitorioya
hoiku	hokennenkin	hokubu	honbun	houkoku	houmon	hozen
html	hyou	hyouka	hyoushi	ichiran	iinkai	iken
img	industry	info	infomation	int	intro	introduction
ishikawa	iso	itiran	jidou	jidouteate	jimu	jinji
jinken	jinzai	jisedai	jisseki	josei	jpki	jyouhou
jyourei	kagaku	kahokugata	kaigi	kaigiroku	kaiho	kaikaku
kaikei	kanazawa	kankyou	kanri	kansa	kansen	kanzai
keiei	keikan	keitai	kekka	kenchiku	kengaku	kenko
kensei	kentiku	kessan	kid	kigyou	kihon	kijyun
kinkyu	kisya	kodomo	koho	kojin	kokuho	kokusai
korea	koseki	kosodate	koubo	kouen	kouhyou	kouiki
kouji	koukai	kourei	kousou	koutsuu	koutu	kouza
kuma	kurashi	kyougikai	kyouiku	kyuyo	law	legal
library	life	link	list	location	machi	magazine
mail	manual	map	mapsearch	master	matsuri	meibo
menu	message	mizu	mokuji	movie	museum	nanbu
nature	navi	nenkin	new	newpage	nintei	nougyou
nourin	nousei	nouyaku	npo	num	odb	osaka
oshirase	other	outline	page	park	passport	pdf
plan	police	pref	privacy	profile	program	project

pubcomme	qanda	quiz	readstep	recycle	rei	reiki
rekishi	renkei	report	result	roudou	roujin	saigai
sake	sakuhin	san	sangyou	sankou	satoyama	schedule
school	search	seido	seikatsu	seinen	sekaiisan	seminar
senkan	senkyo	service	sesaku	setumei	shiken	shikumi
shimin	shinrin	shinsei	shiryou	shisetsu	shishin	shitei
shohicenter	shoubou	shoukai	sigoto	sikin	silver	singikai
sisetu	sisimai	sisyo	site	sitemap	ski	sonota
soshiki	soudan	sougou	soumu	sport	spring	study
sub	suidou	suisan	sukoyaka	summer	support	syakyo
symbol	syo	syokuin	syomu	syougai	syoumei	syurui
taiken	taisaku	tedori	tenken	tetuduki	theme	title
toc	tochi	todoke	todokede	toiawase	tokubetu	tokutei
torikumi	toukei	touroku	town	ugoki	view	vision
volunteer	wakuwaku	walk	watch	winter	word	www
yakuwari	yoko	yosan	yougo	youkou	yousiki	yuki
zaisei	zeimu	zentai				