JAIST Repository
https://dspace.jaist.ac.jp/

Title	アンカーテキストを用いた属性情報の抽出
Author(s)	太田，茂
Citation	
Issue Date	2007-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/3600
Rights	
Description	Supervisor:鳥澤　健太郎，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

アンカーテキストを用いた属性情報の抽出

太田 茂 (510019)

北陸先端科学技術大学院大学 情報科学研究科

2007 年 2 月 8 日

キーワード: アンカーテキスト, k-means, クラスタリング, 属性, サイトマップ.

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

Copyright © 2007 by Shigeru Ohta

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・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.

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った.

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.