Topic Modelを用いた非文字検索システム

電気電子情報工学科　4年　松浦脩司

背景

現代社会においてウェブは欠くことのできない社会基盤であり，ウェブ上に存在する情報が増加するにつれ，その重要性は行政や企業取引，教育などあらゆる分野でますます高まっている．

近年注目されているのが博物館（あるいは美術館など）の収蔵資料をディジタル化し，ウェブ上で公開するという取り組みである．

この取り組みの目的は資料を半永久的に保存し，資料の検索や閲覧を簡単にすることで学術研究や教育に役立てることである．

本研究は福島県只見町の民具資料を対象として，柔軟性とより高度な検索機能を備えた新たな博物館資料の情報検索システムを提案する．

問題と改善点

民具のデータベース化において考えなければならない主な問題として，民具分類に関する問題と民具名称に関する問題がある.

只見町の民具データベースに求められる能力を検討すると，少なくとも次に挙げる事柄が必要であると考えられる.

- 民具の複数の用途の表現が可能である
- 統一名(一般的な名称) と地方名の表現が可能である
- 民具分類の基準の変更に柔軟に対応可能である
- 民具カード上の情報に対する全文検索に対応可能である

民具カード

国指定・重要有形民俗文化財
「会津只見の生産用具と仕事着コレクション」

民俗資料調査カード

通番号 T・A 540 M　分類番号 11101-1

資料名	(地方名) シゴトシ	(標準名)
寄贈・借用年月日	昭和・平成　年　月　日	
寄贈者	(住所)只見町大字布沢字　番地	
所有者	(氏名)	
材質	木綿	
使用年代	明大昭　年頃から　明大昭　年頃まで　現在も使用中	
使用目的	作業の時に着る。	
収蔵場所	明和中学校寄宿舎	
備考	仕事着	
調査年月日	平成 2年 7月 30日	
調査員	目黒鶴吉	

写真・形状・寸法等

身丈69. 袖丈30. 裄 63.

只見町教育委員会

提案

Topic Modelを用いて，只見町のデータをオントロジではなく確率的に独立した単語の集まりに分類し，検索できるシステムを提案する.

Topic Model

トピックモデルとは，文書が複数の潜在的なトピックから確率的に生成されると仮定したモデルである．

文書内の各単語は，あるトピックが持つ確率分布に従って出現すると仮定する．

トピックモデルでは，トピックごとに単語の出現頻度分布を想定することで，トピック間の類似性やその意味を解析できる．

例えば，大量のニュース記事をもとに記事のタグ付けを自動化させるケースを想定した場合，一つの記事に複数のタグを付与できるトピックモデルの方が，より多くのユーザーに興味ある記事を届けることができる．

オントロジ

オントロジとは本来哲学用語で「存在論」を意味する言葉である.

対象世界（知識領域）をある視点で見たときに現われてくる構成要素（概念）を明示的に表現し，それらの関係を体系的に記述したもののこと.

セマンティックウェブでは概念や意味を共有し、コンピュータが文書の意味を理解したり，情報を再利用したりするための基盤機構として構築される語彙のセットをいう.

非文字の再定義

オントロジを使用する非文字データの分類とは，本来，存在論に乗らない非文字という現象を工学的データベースに乗せるようなものである．本来存在論に乗らない非文字の現象とは，単語-単語間の作用という現象である．

【定義】非文字とは概念間の作用である．

このような作用を表現するためには作用を確率論的な確率変数として捉えることが必要である．

作用=非文字と定義すれば，作用を抽出する確率分布を抽出(学習)することが可能になる．即ち，分析ツールとして機械学習ソフトウェアが使える．

トピックモデルによる非文字表現

(1) 非文字テクストを構文解析し名詞を分離して機械学習データとする.

(2) 分離された大量の名詞（単語）を独立事象と見做し，ノンパラメトリックベイズ（LDA）に入力し，単語のクラスターを計算する.

(3) クラスターに於いて出現確率の高い単語を選択し，トピックとする.

(4) このトピックがクラスターに於ける潜在的確率変数に対応する単語であり「非文字ｚ」である.

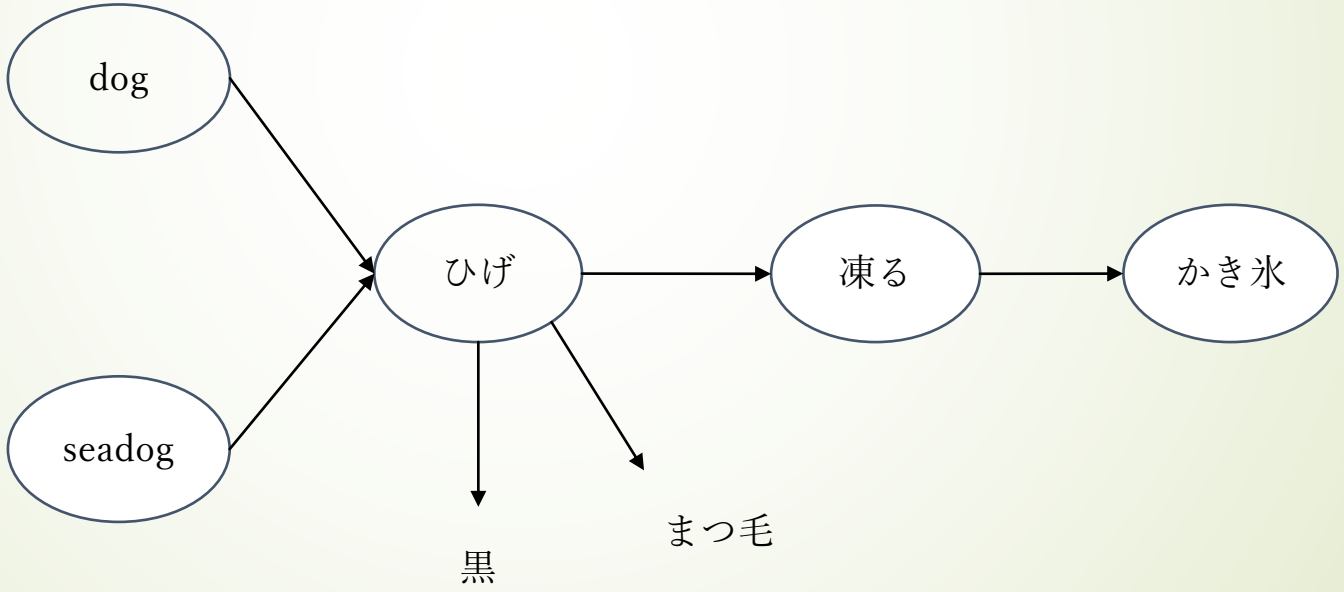
(5) クラスター化された単語の集まりに関与するオントロジξを分析する.

(6) このとき，オントロジξに於ける「非文字ｚ」が求められる

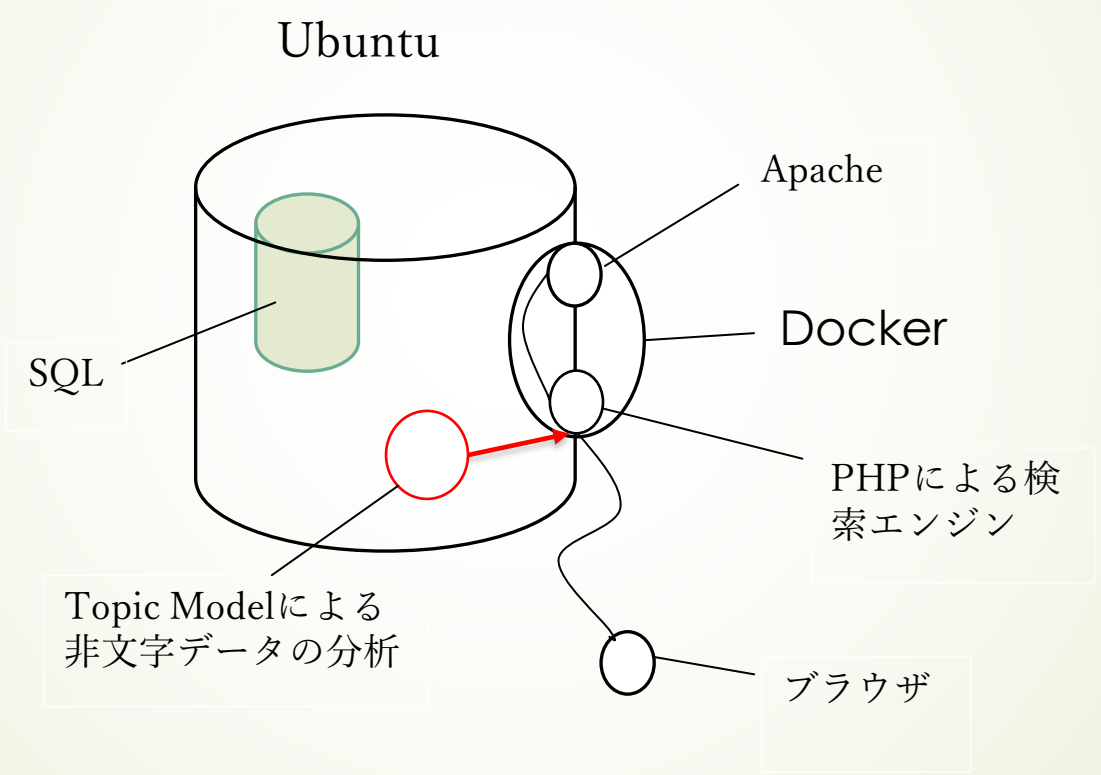
確率分布を抽出するために

- 連想を導く
- この連想を非文字と呼ぶ
- これらを非文字の規則とする
- 個々のコンテナの組み合わせで非文字抽出システムを作りたい
- 開発効率UPのためにDockerを使用する

連想



実装システム


Ubuntu
Apache
Docker
SQL
PHPによる検索エンジン
Topic Modelによる非文字データの分析
ブラウザ

実装システム

実装したシステムを以下に示す.

(1) OS :Ubuntu 18.04

(2) Docker版のApache Web server

(3) Python Anaconda：統合仮想開発環境

AnacondaにはLDAパッケージGensimが実装されており，これを使用すればよい.

結論

トピックモデルを用い，非文字を検索するシステムを提案し実装まで行った.

今後の課題

実装までしかできなかったので，実際に動かしてみて先行研究と比較し，効率の良いシステムか検証する.

もし実際に実装する場合，使いやすいかどうか検証する.