memo

タイトル：
博物館の資料（群）名称の共起関係を利用した統合検索の研究

資料名称の共起関係に基づく連想検索
（同一名称内の共起とコレクション内の共起の区別？）

歴博のデータを対象とした実証的な研究を行う
歴博データ：180087
異なり名称：94109

目的
博物館資料検索において，入力語によるキーワード検索でなく，関連するものをみつけたい
ユーザが正確な資料名称を知っているとは限らない
豊富な対象領域知識も仮定できない

文書検索との相違点
ドキュメント（テキスト）が大量にあるわけではない

メタデータの整備
多大なコスト・手間がかかる
どのようなメタデータをあらかじめ付与しておけば，素人が検索するに足るかが明らかではない

すでにある情報として資料名称，コレクション（一定の資料の集まり）の利用を考える

方法

1．形態素解析
辞書の構築
既存の辞書では資料名称をうまく解析できない
語彙の特殊性
名詞連続，漢字連続，助詞がほとんど用いられない
単語区切り
副次的効果として，読みの付与

2．主要語の選定
長い資料名称の場合，複数の語から構成される
どの語が重要かの選定が必要
予備実験では，関連語として，様々な修飾語が出現

3．共起情報を用いた関連語の計算
コレクションの利用
コレクションに含まれる語
コレクション内の語の共起関係
GETAの利用
類似度の計算方法に関する検討

４．階層性の利用
分類が階層的になされている場合にどのように反映させるか

５．クラスタリング
関連語の体系化

評価用テストセットの作成

問題点
形態素解析の精度→未知語が多い，名詞連続
・ノイズ
チャンキング
・語の粒度（分割しすぎ）
出現した形態素の重み→品詞，ヘッドワードか否かを判定していないため，
共起する語が多く出現
・重要ではない語

コレクションによる違い
（nameとtitleの違い）
古文書の名称をどうするか

・関連性の計算（コレクション，ツリー構造の利用）
・クラスタリング
上位だけしか見ることができない→下位に埋もれた重要語がわからない
すべての候補を見せる方法
見せ方：いくつにわけるか，代表語の選定

利用者の操作・ユーザインタフェース
キーワード入力
〈関連語〉の提示・クラスタリング

手法：言語処理
関連するものを取り出す：連想