資料 7

統合データベース開発：

専門用語辞書管理システムと専門用語解析技術の開発

奈良先端大
松本裕治，新保仁，浅原正幸，
原一夫，鈴木郁美，呂嘉

- 専門用語解析技術
  - 専門用語辞書システムの開発
  - 専門用語解析技術の開発
- 専門用語抽出ツールの設計と開発
  - 専門用語辞書拡張支援ツールの設計と開発

今年度の成果目標

（1）専門用語解析技術

①専門用語辞書システムの開発

- 9月末まで：
  - 10万語以上の規模の生命科学用語の辞書を格納（ライフサイエンス辞書，病名マスターなど）
  - Webブラウザ上で用語の検索や用語のもつシソーラスコードなどの情報を表示・修正機能
- 22年度末：
  - 用語間の意味関係（シソーラスに基づく概念的な上位下位関係や類似度）の表示機能
  - 全体的な専門用語辞書システムを完成させる。

今年度の成果目標

（1）専門用語解析技術

②専門用語解析技術の開発

- 9月末まで
  - 学習データとして内部構造解析済みのデータを2000語以上に拡大
- 22年度末：
  - 前年度に設計した一般的な統語解析法に基づくアルゴリズムをコンピュータ上に実装し、90%以上の内部構造解析精度を達成する

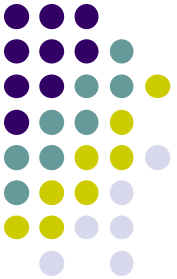
今年度の成果目標

(2)専門用語抽出ツールの設計と開発

①専門用語辞書拡張支援ツールの設計と開発

平成21年度までに、専門分野のテキストに現れる新規の専門用語（新規語）と既知語との意味的な類似度計算する手法を実装し、新規語と意味的に類似性の高い専門用語辞書（シソーラス）内の既知語候補を，シソーラスの構造とともに提示するインタフェースの設計を行った。

- 今年度9月末まで:
  - 用語の内部構造の情報など，種々の情報を用いて類似度判定の精度を向上を図る
  - 新規語に対してシソーラスコードを付与するための支援機能
- 22年度末:
  - 専門用語辞書システムの一機能としてWebブラウザ上で利用可能な機能として統合する。



専門用語解析システム
1．専門用語辞書システム

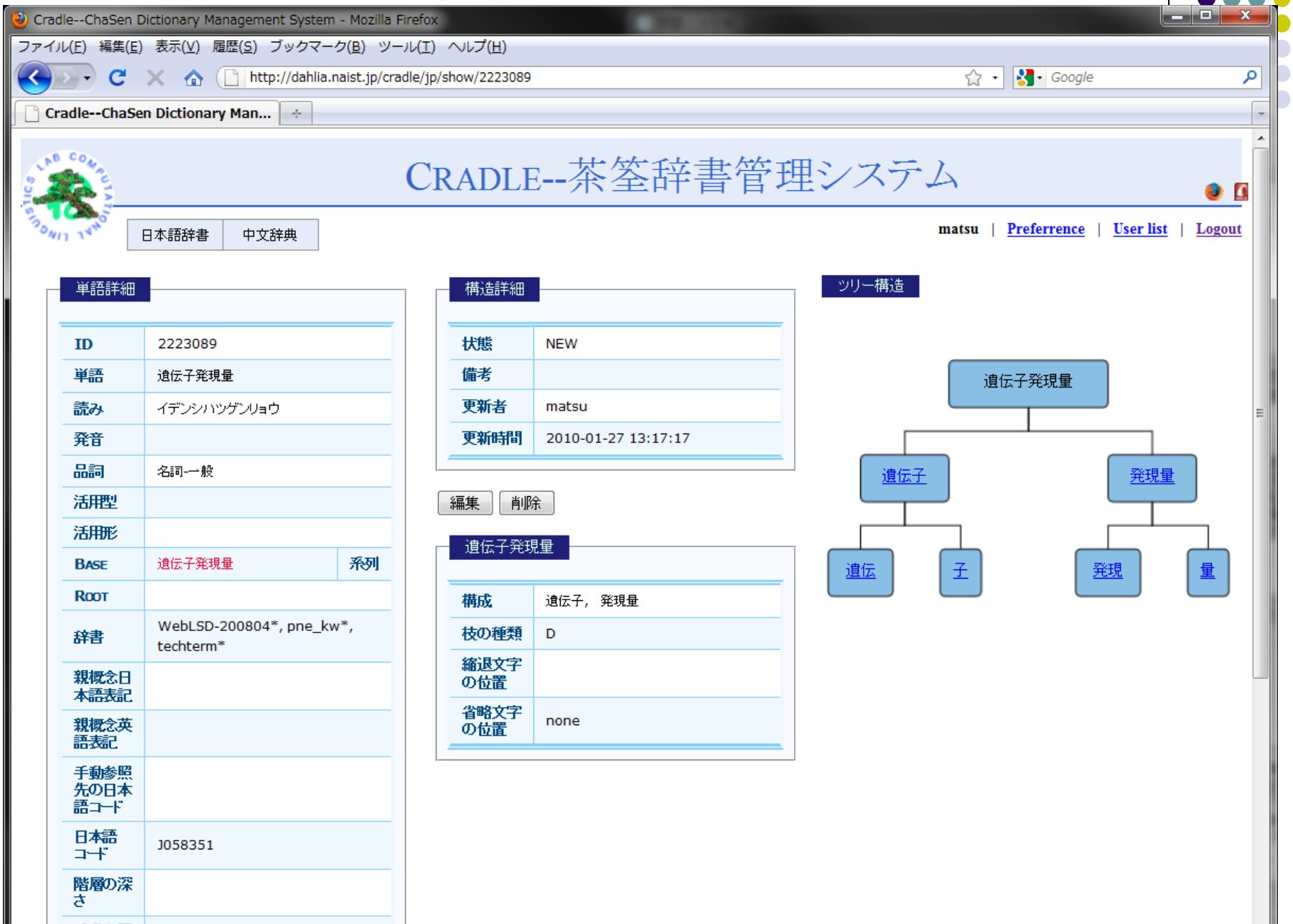
辞書管理システムCradle

- 形態素解析用辞書の管理ツール（現状）
  - 現在登録している辞書，用語集
    - ライフサイエンス辞書（京大金子研究室）
    - 標準病名マスターv2.80を格納
    - 仲里さんよりいただいた専門語候補(18万語)
  - 複合語に対する内部構造付与
    - 現在約1800語について人手により内部構造付与
  - 辞書の標準項目による検索以外に，同義語検索，内部構造に基づく検索を実装
  - 表示項目のカスタマイズ機能を実装をより一般化

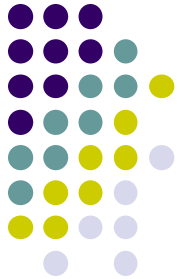
検索画面


Cradle--ChaSen Dictionary Management System - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://dahlia.naist.jp/cradle/
Google
Cradle--ChaSen Dictionary Man...
CRADLE--茶筌辞書管理システム
日本語辞書
中文辞典
matsu | Preferrence | User list | Logout
単語属性
リセット
検索
ID
単語
読み
発音
品詞
活用型
活用形
Base
辞書
NAIST-jdic-20080707
WebLSD-200804*
標準病名マスター V2.80*
pne_kw*
techterm*
文字数
状態
新規者
更新時間
更新者
親概念日本語表記
親概念英語表記
手動参照先の日本語コード
日本語コード
階層の深さ
手動参照先の日本語表記
自動参照先ID
ツリー番地
自動参照先表記
ツリー日本語
親概念ID
ツリー英語
ICD10
頻度(文数)
確信度
頻度(文献数)
複合語属性
リセット
検索
内部表記
include
内部読み
include
内部POS
include
状態

単語情報の表示

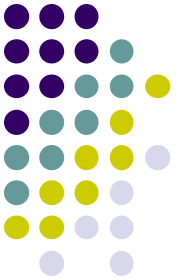

Cradle--ChaSen Dictionary Management System - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://dahlia.naist.jp/cradle/jp/show/2223089
Google
Cradle--ChaSen Dictionary Man...
CRADLE--茶筌辞書管理システム
日本語辞書
中文辞典
matsu | Preferrence | User list | Logout
単語詳細
ID 2223089
単語 遺伝子発現量
読み イデンシハツゲンリョウ
発音
品詞 名詞一般
活用型
活用形
BASE 遺伝子発現量 系列
ROOT
辞書 WebLSD-200804*, pne_kw*, techterm*
親概念日本語表記
親概念英語表記
手動参照先の日本語コード
日本語コード J058351
階層の深さ
構造詳細
状態 NEW
備考
更新者 matsu
更新時間 2010-01-27 13:17:17
編集
削除
遺伝子発現量
構成 遺伝子, 発現量
枝の種類 D
縮退文字の位置
省略文字の位置 none
ツリー構造
遺伝子発現量
遺伝子
発現量
遺伝
子
発現
量

専門用語辞書システムの開発項目



- 辞書システムの機能拡張
  - 表示項目の追加･削除，値の属性定義のカスタマイズ
    - 辞書によって表示項目を自由に定義できるようにする
    - DBシステムをMySQLからMongoDBへ変更予定
  - 管理者権限の分類，グループ化
    - 一部のDBにのみ権限をもつ管理者
    - 一部のDB修正機能にのみ権限をもつ管理者

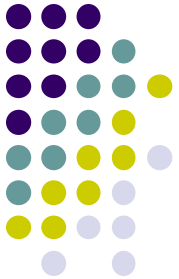
    など，最上位の管理者以下に様々な異なる管理者グループを定義できるようにする



専門用語解析技術
２．専門用語解析技術の開発

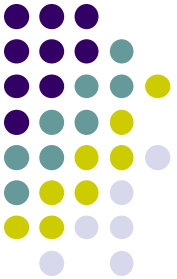
専門用語解析技術の開発項目

- 内部構造解析データの拡張
  - 現状の1800語を2000語以上に拡張
    - できれば3000語以上の用語の内部構造タグ付けを行いたい
  - 内部構造の自動解析システムの実装・評価
    - 昨年度までに設計した統語解析アルゴリズムの一般化に基づく方法を実装
    - 半教師付き手法による精度向上実験
    - これまでの実装（文字ベースの決定性アルゴリズム）との比較評価



（2）専門用語抽出ツールの設計と開発
専門用語辞書拡張支援ツール

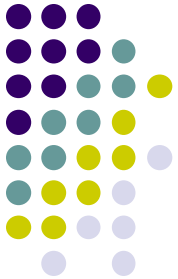
研究の目的＝シソーラス拡張



- 新規の専門用語に対して，類似度が高い順に登録済の専門用語をランク付けし，提示するシステムの構築

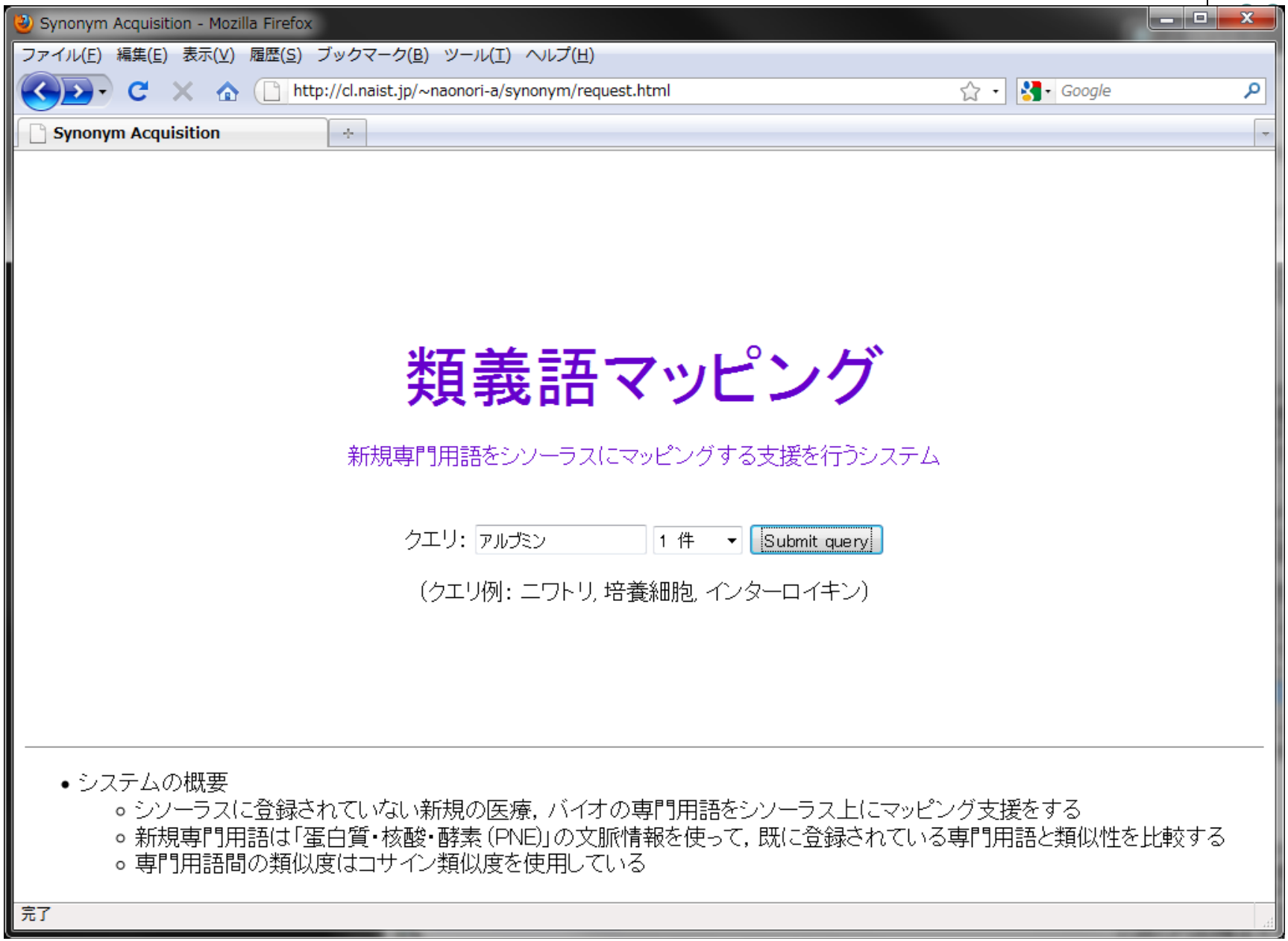
- シソーラス辞書の編集者は，システムが提示するランキング上位語を参考に，新しい専門用語をシソーラス辞書に格納する

専門用語辞書拡張支援ツール

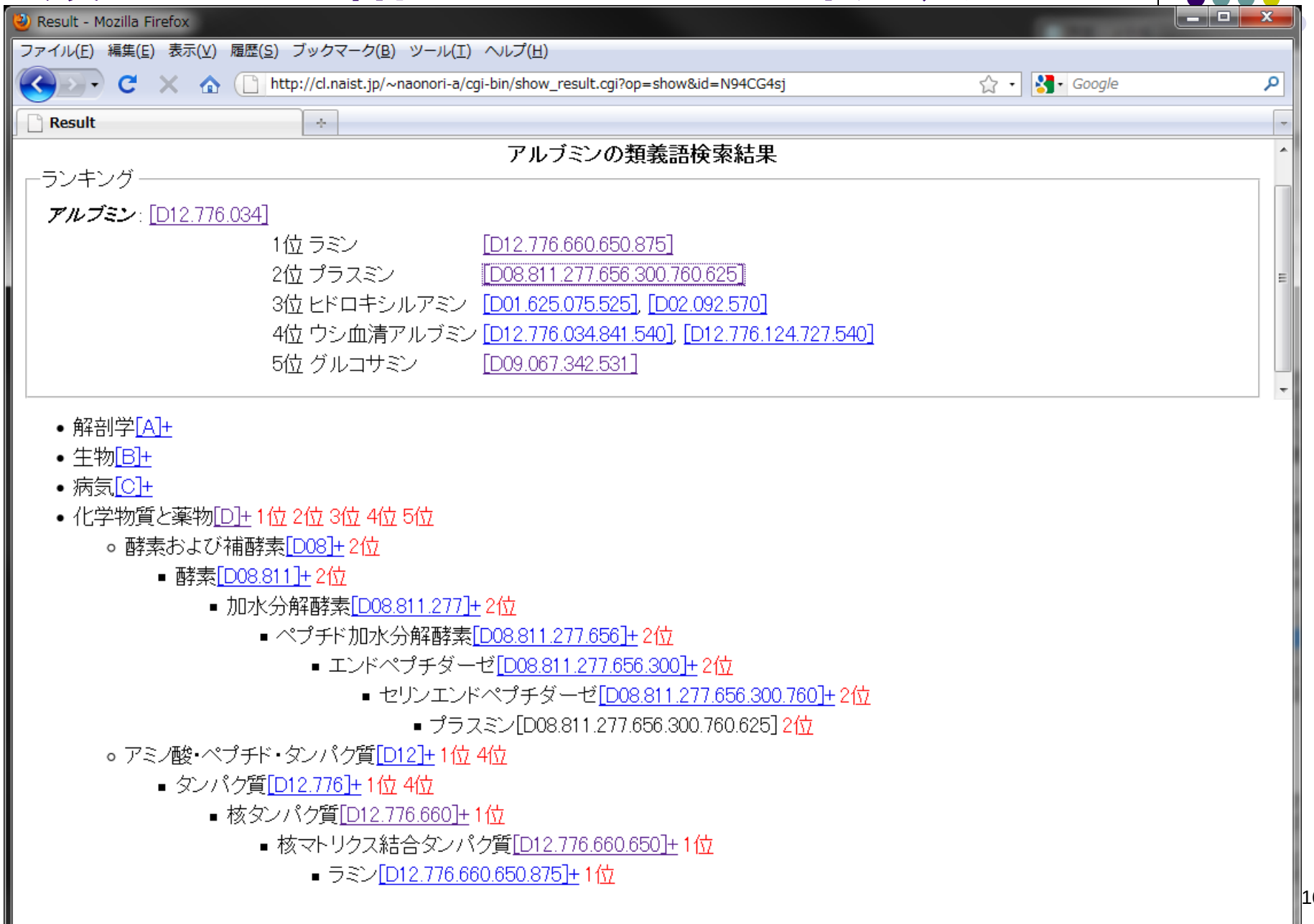


- 専門文書から対象とする語の文脈情報を抽出して用語の隣接グラフを作成し，グラフ構造を用いて用語間の類似度を算出する手法を提案
- 雑誌「蛋白質・核酸・酵素」を実験データとして用い，そこに登場する新規の専門用語と類似度の高い語をライフサイエンス辞書から検索する
- 検索した類似度上位の語のシソーラス内の位置を表示するインタフェースを構築

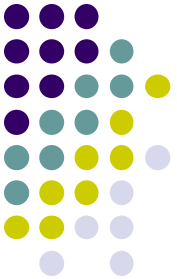
類義語検索ツールの初期画面



類似の上位語とシソーラスの部分表示



専門用語辞書拡張支援ツールの今後の予定



- 9月末まで:
  - 用語の内部構造の情報などの情報も利用することにより類似度判定の精度を向上
  - 新規語に対してシソーラスコードを付与するための支援機能
- 今年度末まで:
  - 専門用語辞書システムと連携させ，辞書システムから呼び出すことにより，辞書システムの一機能として利用可能にする