青山学院大学　理工学部　情報テクノロジー学科

平成１７年度卒業論文要旨

意味グラフを利用した全文検索システムの開発

川又 真綱（15802026）小松 勇也（15802034）近藤 哲司（15802035）
原田研究室

１．はじめに

従来の検索システムはキーワードを複数個指定し，それらを含む文章を検索するものが多かった．これでは，キーワードと同意だが表層的に違う単語で記述されている記事を検索できない．また，キーワード同士の関係を指定できないため、検索主旨に合わない文章を多く検索してしまう．これらの問題点を解決するために，キーワードの代わりに検索文を用い，検索文と検索対象となる知識文章の意味的な類似度を計算することで，有用な情報を検索できるシステムを構築する．

２．全文検索システム Reans

Reans は，日本語で書かれた検索文と、知識文章全体の文章類似度を計算することで，ユーザの要求する知識文章を出力する検索システムである．知識文章のデータベースである ReansDB と意味検索システム ReansIR の 2 つから構成される．図１に Reans の全体図を示す．

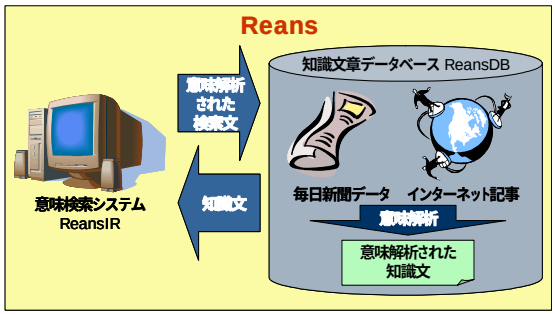

図 1 Reans の全体図

２．１．知識文章データベース ReansDB

ReansDB には，毎日新聞記事 4 年分やインターネットから集めた 73 万件の記事データと，それらをあらかじめ日本語意味解析システム SAGE により意味解析をした意味グラフを蓄える．両データはファイル名で対応付ける．

２．２．意味検索システム ReansIR

ReansIR は下記の手順で検索趣旨にあう知識文章を検索する．

①検索文の意味解析：検索文を SAGE によって意味解析し検索グラフとする．②キーワード抽出：SVM を用いて，検索グラフからキーワードを抽出し，非線形判別式の得点順にソートしたものをリストにする．③キーワードの同意語展開：各キーワードの概念 ID から同意語を調べキーワードに含める．例えば「事業」と「ビジネス」のような表現の揺れを吸収することができ，より多くの知識文章を検索できる．④全文検索：拡張されたキーワードを基に Namazu を用いて検索を行い，得られた知識文章の知識グラフのファイルパスを得る．⑤文章類似度の計算：検索グラフと得られた知識文章全ての知識グラフを照合し，その間のグラフ類似度を用いて文章類似度の算出を行う．⑥結果の出力：文章類似度の降順にソートして結果として表示する．図２に結果の出力例を示す．

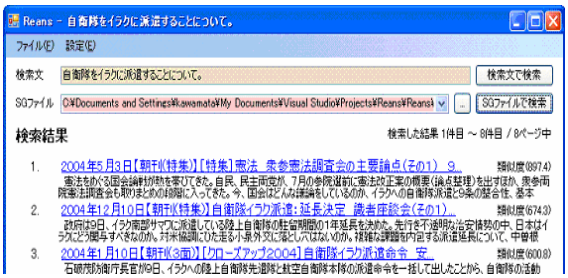

図 2　結果の出力例

３．おわりに

本システムを利用した検索と，Namazu 検索の結果を比較すると，Namazu のみの検索では検索趣旨と異なる文章が上位に表示されたが，本システムを利用した検索では，検索主旨に沿った文章を上位に表示させることができた．よって，本システムの有効性を示せたと言える．