2C-6

# 検索文と記事文章全体との照合による意味検索システム Reans

河村 惇史† 久保谷 篤† 松田 源立‡ 原田 実‡
青山学院大学理工学部情報テクノロジー学科†,‡

## 1. 背景

今日我々はインターネット等を介し，大量かつ様々な情報を得ることができる．従来の検索システムでは，キーワードの関係の指定が不可能であり，ユーザの意図に沿わない文章が検索されてしまうことがある．また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．そのため得られる文章は，要求に合わないことが多い．

## 2. 目的

本研究では以上のような問題点を解決するために，意味検索システム Reans を構築する．Reans は，キーワードの代わりに検索文を用い，検索文を意味解析して得た検索グラフと，予め意味解析した知識グラフ群との文章適合度を算出し，これを元に知識文書をランキングする．そのために，学習を用いて良い得点を出すパラメータを得る．それによりユーザの質問意図に即した検索結果を得ることができる．

## 3. 意味検索システム Reans

Reans は図 1 のように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．

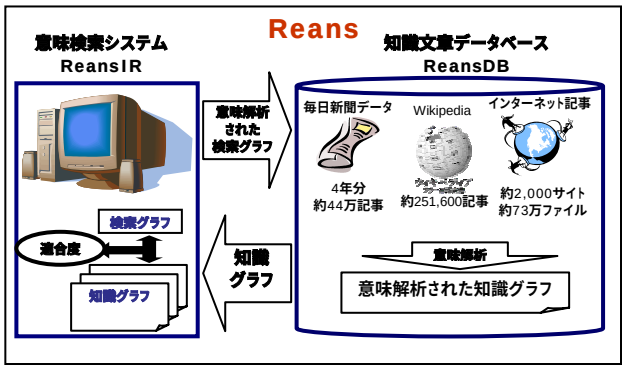


図 1 Reans の構成

### 3.1. ReansDB

ReansDB には現在，毎日新聞記事 4 年分，インターネットから集めた 73 万件の記事データ，Wikipedia[1]の記事約 25 万記事を，日本語意味解析システム Sage[2]によって意味解析した知識文(知識グラフ)が含まれている．

### 3.2. ReansIR

ReansIR の検索方法は図 2 のようになっている．

知識文を探索するには以下の手順で行う．検索文を日本語意味解析 Sage で意味解析し検索グラフとする．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する．また各キーワードの語彙から同意語を調べキーワードに含める．これにより例えば「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の検索が可能になっている．その後全文検索システム Lucene[3]を利用し，検索結果を得る．

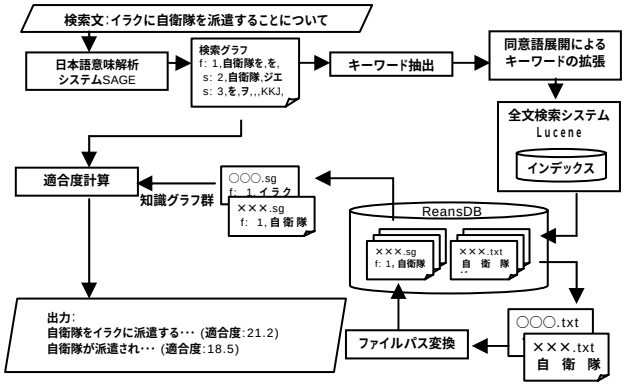


図 2 ReansIR の構成

#### 3.2.1. 文章適合度の計算

前のステップでキーワードをすべて含む文章を取得したので，次に，取得した文章を文章適合度の高い順番に並び替える．このために，取得した知識文章の各文と検索文との文類似度や，キーワードの位置による類似度などの計算を行う．

文章適合度は、これらの 11 個の適合度得点$X_i$に，重み$W_i$を掛けたものの合計とする．

$$\text{文章適合度} = \Sigma \quad W_i X_i \qquad (3)$$

適合度得点$X_i$は，文類似度やキーワードの出現位置による得点を用いて作成した．

その適合度得点は，検索文と知識文との補正文類似度の最大($X_1$),検索文とタイトルとの文類似度($X_2$)，最初の 3 文の補正文類似度の合計($X_3$),タイトルに含まれるキーワード数($X_4$),1 文に含まれる平均キーワード数($X_5$)，1 文に含まれる最大キーワード数($X_6$)，3 文に含まれる最大キーワード数($X_7$)，最初の 3 文に含まれるキーワード数($X_8$)，すべてのAND検索キーワードが含まれている文の数($X_9$)，同意語展開前のキーワードが使われている割合($X_{10}$)，Luceneのスコア($X_{11}$)である．

文類似度は，意味解析された知識グラフと検索文の意味グラフの類似度であり，下記の式を用いて求める．

$$\text{文類似度} = (1-\alpha)\times\text{ノード類似度} + \alpha\times\text{アーク類似度} \qquad (1)$$

α：関係重視率（0≦α≦1）

$$\text{ノード類似度} = \frac{\Sigma\text{照合ペア概念類似度}}{\text{グラフのノード数}}$$

---

**A semantic retrieval system Reans based on semantic matching between retrieval sentence and the entire sentences of article**

Atsushi Kawamura†, Atsushi Kuboya†, Yoshitatsu Matsuda‡ and Minoru Harada‡
†Undergraduate school of Integrated Information Technology, Aoyama Gakuin University.
‡Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University.

$$\text{アーク類似度} = \frac{\Sigma(\text{深層格類似度+アーク両端のノード類似度})}{\text{グラフのアーク数}}$$

ただ，この文類似度は意味グラフの構造をみているので，検索文と知識文の長さが異なる場合，同じことを述べていても類似度があまり上がらないことがあるので，下記のように補正を行う．

補正文類似度 ＝

文類似度 × 知識文に含まれるキーワード数 × 0.5 (2)

この式により，知識文にキーワードが 2 つより多く含まれている場合，文類似度は大きく，少ない場合小さく補正される．

### 3.2.2. 遺伝的アルゴリズムによる重みの決定

ReansIRは文章適合度に基づき知識文章を並べている．その並びが手動で作成した正解データとどの位一致しているかを，スピアマンの順位相関係数を用いて評価する．この順位相関係数が最大になるような$W_i$を決定する必要がある．そのために，順位相関係数は微分不可能であることから，遺伝的アルゴリズムを利用した．

1 個体の遺伝子の並びを用いて$W_i$を求める．図 3 に 1 個体の遺伝子の詳細を示す．

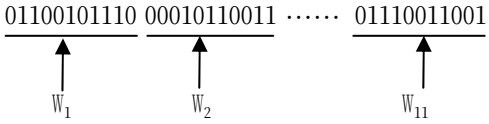


図 3　1 個体の遺伝子

$W_i$は 0 から 1 までのおよそ 0.001 刻みの値を取るようにする．そのために，最大 1023 の値を取る 10 ビットに，グレーコードのための 1 ビットを付加した 11 ビット与えている．このビット列をバイナリーコードにした後，10 進数に直す．この値を得られる最大値 1023 で割ることで 0 から 1 の$W_i$の値を求める．この$W_i$を元に(3)式を用いて文章適合度を求め，その値の高い順に並び替える．この順位と，手作業で作成した正解順位との順位相関係数を求め適応度とする．遺伝子の選択にはルーレット選択方式を用いる．個体iの適応度$f_i$を元に，選択する確率$P_i$を以下の式で求めることができる．

$$P_i = \frac{f_i}{\sum_{j=1}^{n} f_j} \qquad (4)$$

この式で求めた確率に基づき個体を選択する．選択された遺伝子を 2 点で交叉をし，この遺伝子に，確率 1%で突然変異を発生させ，次の世代の遺伝子を作成する．

ただ，ルーレット選択方式では確率的に次世代の遺伝子群を決定するため，最良の個体が次世代に残らない可能性がある．これを避けるため，エリート選択方式を使い，適応度が高い遺伝子 m 個を次の世代に残し，残りをルーレット選択方式により選択する．

この操作を繰り返し，最も順位相関係数の高くなる$W_i$を得る．

## 4. 評価実験

実験データとして使用したのは，2004 年度の毎日新聞記事データである．

まず，100 個の検索文を作成し，それによってヒットした合計 1666 の知識文章を用いて実験を行う．

遺伝的アルゴリズムの適応度に順位相関係数を用いているが，そのためには正解データとなる正しい順位を作成する必要がある．検索文で得られた知識文章すべてを，手作業で 1 から 5 のランク付けを行った．このランクを元に知識文章の順位を決定した．ランクが同じものがある場合はそれを同順位とした．これを 100 個の検索文すべてで行い，学習のための正解データとした．

遺伝子の数を 100 個，エリート選択によって次の世代に残す遺伝子を 5 個，世代数を 2000 世代として実験を行った．その結果，得られた重みの値$W_i$と，適合度得点Xiの平均値は以下の表通りである．

表 1　$X_i$の平均値と重み$W_i$の値

| i | $X_i$の平均値 | $W_i$ | $X_iW_i$の平均値 |
|---|---|---|---|
| 1 | 39.2 | 0.00684 | 0.268 |
| 2 | 4.00 | 0.0420 | 0.168 |
| 3 | 26.0 | 0.0410 | 1.07 |
| 4 | 1.59 | 0.342 | 0.544 |
| 5 | 0.394 | 0.603 | 0.238 |
| 6 | 3.98 | 0.546 | 2.17 |
| 7 | 5.73 | 0.288 | 1.65 |
| 8 | 3.99 | 0.931 | 3.71 |
| 9 | 0.357 | 0.137 | 0.0489 |
| 10 | 0.891 | 0.877 | 0.781 |
| 11 | 0.320 | 0.768 | 0.246 |

表 1 より$X_8$の最初の 3 文に含まれるキーワード数が最も重要視されていることがわかる．

このときの順位相関係数は，0.510 である．すべて正解の通りに並べた場合の順位相関係数は 0.826 で，Lucene 検索のみの場合は 0.362 である．

Lucene は TF-IDF 法でスコアを算出し，順位を決めている．本システムを利用した検索と，Lucene 検索の順位相関係数を比較すると，Lucene 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．

## 5. まとめ

意味検索システム Reans は，SAGE を使っているので，一般的な検索システムとは異なり，調べたいことを文で入力することができる．これにより，ユーザは検索キーワードの選定をしなくてよいため，より感覚的に調べることができる．また，検索文と知識文章との文章適合度を用いることで，ユーザの意図に合った文章の順に並び替えることができるようになった．

現在，実験に用いたのは毎日新聞記事データのみなので，今後は，インターネット記事や Wikipedia に対応した，適合度得点の選定や，重み付けを行っていきたいと考えている．

## 参考文献

[1] フリー百科事典 Wikipedia
http://ja.wikipedia.org/
[2] 杉村和徳,山本哲哉,木村健太郎,鳥居隼,韓東力,原田実:”意味解析システム SAGE の精度向上と利便性の向上”, 情報処理学会第 67 回全国大会論文集,1J-02,第 2 分冊, pp.67-68 (2005.3).
[3] 全文検索システム Lucene :
http://lucene.apache.org/java/docs/