3D1-3

概念グラフマッチングによる自然言語テキストの意味検索

Semantic retrieval for natural language texts by using conceptual graphs

高山 智史[*1] Satoshi Takayama
石塚 満[*1] Mitsuru Ishizuka
内田 裕士[*2] Hiroshi Uchida

[*1] 東京大学大学院情報理工学系研究科 Graduate School of Information Science and Technology, The University of Tokyo
[*2] UNDL 財団 UNDL Foundation

In this paper, we propose a semantic retrieval system using conceptual graphs. Our system processes natural language texts as CDL (Concept Description Language) and extends queries with Universal Word thesaurus. In our experiments we evaluate two aspects of our system: query extension and use of conceptual relations. The results showed that our methods improved the retrieval performance.

1. はじめに

人類がアクセス可能な情報の量はインターネットの普及によって爆発的に増加しており，それらの膨大な情報の中から必要とする情報を探し出す方法として検索エンジンが重要な役割を担っている．しかし，要求に適合する情報を十分に絞り込むには現状の検索システムでは不十分であると考える．例えば，従来のキーワードによるテキスト検索では，キーワードを含むテキストを探し出すことはできるが，その内容が要求に適合しているかどうかは人間が読んで判断しなければならない．このような問題を解決するためには，将来的にテキストの意味を考慮した検索システムの実現が不可欠であると考える．

本稿では，意味を考慮したテキスト検索として CDL によるテキスト検索システムを提案する．CDL を用いることで，語と語の間の関係など柔軟な検索条件を指定でき，また，人間によるテキスト理解を経ずに直接コンピュータが CDL を扱うことで，より高度な情報処理が可能になると考える．

本稿では以下 2 章で検索システムの概要について説明し，3 章でシステムを用いた実験について述べ，4 章でまとめと今後について言及する．

2. システム概要

2.1 CDL

我々のシステムでは自然言語テキストを Concept Description Language (CDL) [石塚 06] として扱う．CDL では実体(Entity)を表すノードと，Entity 間の関係(Relation)を表すエッジからなるグラフ構造によって文章を表す．以降，Entity-Relation-Entity の 3 つ組を Triple，CDL 化された文章を CDL 文と呼ぶ． CDL 文の例を図 1 に示す．

例文: John bought a computer yesterday.

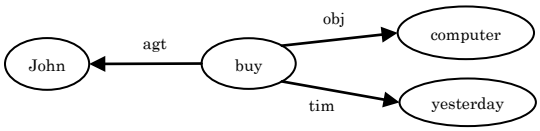

図 1: CDL 文の例

連絡先：高山智史，東京大学大学院情報理工学系研究科，
takayama@mi.ci.i.u-tokyo.ac.jp

2.2 Universal Word

本研究では CDL で利用する語彙として，UNDL によって開発された Universal Word (UW) [Uchida 05] を利用している．また，クエリ拡張で利用するために UW の概念階層構造を元にシソーラスを構築し，更にクエリ拡張を効率的に行うために階層的なコーディングを施した値(UWCode)を独自に UW に付与している．UW の概念階層の一部を図 2 に示す．

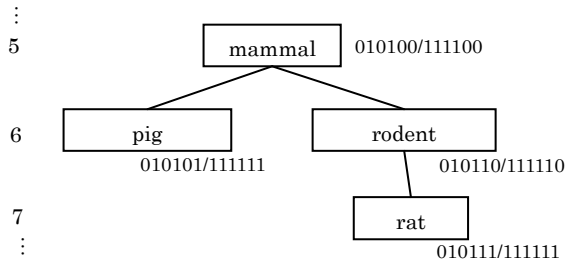

図 2: UW の概念階層と UWCode の例

2.3 CDL の検索

CDL 文の検索処理とは，基本的にクエリとして CDL 文の部分グラフ(を構成する Triple)を与え，その部分グラフを含む CDL 文を返すというものである．ただし，部分グラフの各ノードはシソーラスを用いて上位・下位概念でのマッチも許可するように拡張される．

3. 実験

実装した検索システムを利用し，以下の 2 つの実験を行った．

- 2 種類のクエリ拡張法での応答時間を比較する実験
- Entity 間の意味的役割を用いた場合と用いない場合とで精度を比較する実験

3.1 実験データ

検索対象データとして Wikipedia の一部のページを CDL 化したデータを使用した．使用した CDL データの情報を表 1 に示す．

検索に用いるクエリについては，以下の 2 種類のクエリ集合を用意した．

表 1: CDL データの情報

CDL 文の数	2770
Triple の数	25115
ユニークな UW の数	7390

(1) クエリ集合 1

CDL データから CDL 文を 10 個選び, 各 CDL 文から 3 個の Triple を選んで Triple3 個 1 組からなるクエリを 10 個作成した.

(2) クエリ集合 2

クエリ集合 1 の各クエリについて, いくつかの Entity を上位概念・下位概念に置換したクエリを 10 個作成した.

Triple3 個 1 組のクエリの例を図 3 に示す.

[agt, expose, server]
[obj, expose, system]
[mod, system, data]

図 3: Triple 3 個 1 組のクエリの例

3.2 実験 1

効率的なクエリ拡張方法を検討するために, 以下の 2 つの方法でクエリ拡張を行い, 検索の応答時間を測定し比較した.

(1) 方法 1

クエリ中の各 Entity に対し, シソーラスの上位 2 階層までの上位概念, 下位 3 階層までの下位概念の UW を抽出し, それらの UWCode を OR で結合したクエリを用いる.

(2) 方法 2

方法 1 において, 下位概念の UWCode を OR で結合して羅列する代わりに, UWCode の最小値, 最大値を用いて範囲で検索するクエリを用いる.

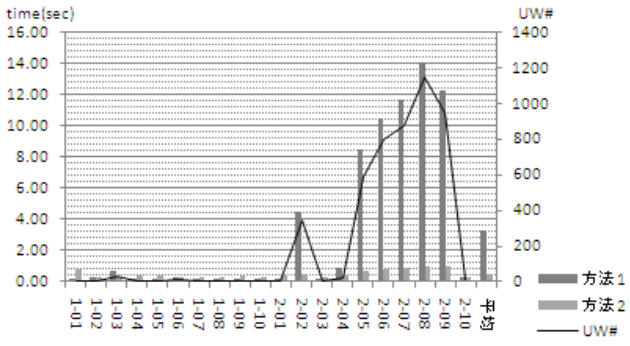

図 4: 応答時間

応答時間の測定結果を図 4 に示す. 横軸の 1-01～1-10 はクエリ集合 1, 2-01～2-10 はクエリ集合 2 での結果である. 方法 1 では, 特にクエリ集合 2 でクエリ拡張による UW 数が多くなるほど応答が遅くなり, 方法 2 では応答時間を短縮できていることがわかる.

3.3 実験 2

検索に Entity 間の意味的役割を用いることの効果を確認するために, クエリに Relation を用いた場合と用いない場合とで検索結果の平均精度を測定した.

検索結果のランキングに用いるスコアの計算には [Zhong 02] の概念間距離を使用した. 具体的には, クエリ Q を構成する Triple の集合を $T_0 = \{T_{01}, T_{02}, \cdots\}$, T_{0j} を構成する Entity の UW を U_{0j1}, U_{0j2} とし, Q に対する検索結果の集合を $R = \{R_1, R_2, \cdots\}$, R_i を構成する Triple の集合を $T_i = \{T_{i1}, T_{i2}, \cdots\}$, T_{ij} を構成する 2 つの Entity の UW を U_{ij1}, U_{ij2}, U のシソーラス上での概念階層の深さを $d(U)$ とすると, 2 つの UW の概念間距離 $D(U_{0jk}, U_{ijk})$, Triple T_{0j} と T_{ij} との間の類似度 $S(T_{0j}, T_{ij})$, 検索結果 R_i のスコア $Score(R_i)$ は以下の式で計算される.

$$D(U_{0jk}, U_{ijk}) = |2^{-d(U_{0jk})} - 2^{-d(U_{ijk})}|$$

$$S(T_{0j}, T_{ij}) = 1 - D(U_{0j1}, U_{ij1}) - D(U_{0j2}, U_{ij2})$$

$$Score(R_i) = \sum_j S(T_{0j}, T_{ij}) / |T|$$

20 個のクエリに対する検索結果のスコアから MAP(Mean Average Precision) を求めたところ, Relation を用いない場合は 0.768, Relation を用いた場合は 0.843 であり, Relation を用いることの有効性が確認できた.

4. おわりに

本稿では CDL による意味的検索システムを提案し, また, 実際に作成したシステムを用いた 2 つの実験について述べた. クエリ拡張を用いた場合の応答時間を測定する実験では, UWCode の範囲を用いた手法で応答時間を短縮できることが確認できた. クエリに Relation を用いた場合と用いない場合とでの精度を比較する実験では, Relation 使用の有効性が確認できた.

今後は, システムのさらなる性能向上のためのクエリ拡張法の検討や, 大規模なデータに対する実験を行う予定である.

参考文献

[石塚 06] 石塚満: 自然言語テキストの共通的概念記述, 人工知能学会誌, Vol.21, No.6, pp.691-698 2006.11

[Uchida 05] Uchida, H., Zhu, M.: from Language Infrastructure Toward Knowledge Infrastructure. http://www.undl.org/materials/UNL2005.pdf (Retrieved April 2010)

[Zhong 02] Jiwei Zhong, Haiping Zhu, Jianming Li and Yong Yu: Conceptual Graph Matching for Semantic Search, Proc. of 10th International Conference on Conceptual Structures, 2002.