概念記述言語CDLに対する意味的検索手法の高度化

西野兼治　堀内暢之　石塚満

東京大学工学部電子情報工学科

1. はじめに

コンピュータ上で扱われる情報には自然言語をはじめとして多様な表現方法があるが、コンピュータはそこから意味を抽出することはできない。そこで様々な表現方法、メディア、コンテンツの意味を共通した形式で表現することでコンピュータに意味を扱わせようとする試みがあるが、その一つにセマンティック・コンピューティング研究開発機構(ISeC)により策定された概念記述言語CDL(Concept Description Language)がある。これによって記述されたデータに検索を行うことで意味的な検索が可能となり[石塚 2009]、そのような検索手法としてSQLを用いた方法[堀内 2012]が提案されている。しかし、この方法は検索条件に厳密に一致した情報を検索するには有効である一方、一致はしないが意味的に似ている情報も得たいという場合には不十分なものであった。本研究では実行するSQL文を変更し、検索条件に意味的に近い情報を検索する手法を提案する。

2. CDL

CDLは自然言語などで記述された多様な意味内容を表現するための言語であり、そのデータは概念を示すentityをノードとし、その概念間の関係を示すrelationをアークとした有向ハイパーグラフで表される[ISeC 2007]。CDL文の例を図1に示す。

例文 :Alice bought pencils.

<#0:buy @ past>
<#1:Alice>
<#2:pencil @ pl>
[#0 agt #1]
[#0 obj #2]

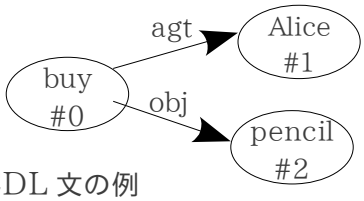

図1: CDL文の例

3. 検索システム

3.1 先行研究

CDLデータをSQL上に保存し、CDLに対する検索クエリとしてCDQLを用いて検索を行うシステムが堀内氏によって提案されている[堀内 2012]。これによってCDLデータの中からクエリのグラフに一致する構造を持つ部分グラフを包含するentityを検索できる。CDQLクエリの例を図2に示す。

```
GET ENTITY WHERE
{?
   <?07:cure+>
   <?0O:end>
   <?27:likely>
   [?0O aoj ?07]
   {?01
      <?0X:medical condition>
   }
}
```

図2:CDQLクエリの例

このシステムでは、CDLデータのentity, relationをそれぞれentityテーブル、relationテーブルに保存し、検索の際にはクエリのグラフ構造に合わせて各テーブルをINNER JOINによって再帰的に結合していく。この際に、クエリの構造に一致しないentityには結合する行が存在しないため、これらは結果となるテーブルから取り除かれる。次にこうして得られた結果テーブルに対し、entityのノードの内容を示すそれぞれの定義ラベルをクエリのものと比較し、一致しない行を取り除く。以上のようにして同じ構造、同じ内容のentityを示す行のみの集合が結果テーブルとして得られる。

しかしながら、この手法ではクエリのグラフと同一ではないが意味的に近いグラフを発見することは難しく、求める情報に対して正確なクエリが要求されると

いう問題点があった。そこで定義ラベルの比較において、条件を一致だけではなく、クエリのラベルの下位語にあたるものも認めるとする改善が行われているが、その他の改善は行われていなかった。

3.2 提案手法

意味的に近い検索を可能とするため、前述のシステムを以下のように変更する。

3.2.1 定義ラベルの条件緩和

定義ラベルを比較する際に上位語に加え下位語、同族語の場合も結果として認める。これにはWordNet3.0 のオントロジを用い、品詞情報などは無視して拡張を行った。

3.2.2 relation の条件緩和

定義ラベルの条件緩和と同様に relation の種類を比較する際に、クエリのものとの類似度が 0 でない relation も結果として認める。なおこのとき用いる類似度はデータベースから得ており、容易に変更することができる。

3.2.3 クエリの部分グラフによる検索

既存のシステムでは INNER JOIN の際にクエリの構造に一致しない entity が結果から取り除かれていた。この方法では、クエリの一部が欠けたような部分グラフと一致するグラフが CDL データベースにあった場合にも、それを結果として認めず、図 3 に示したようにクエリと意味的に似ていても除外してしまう問題点があった。

そこで、クエリの部分グラフでも検索を行えるようにするために INNER JOIN ではなく LEFT OUTER JOIN を用いてテーブルを結合するものとした。ここで、全ての部分グラフを検索しようとすると時間がかかりすぎてしまうため、結果のグラフが含んでいるべき entity をクエリのグラフから一つ指定し、その entity を含む部分グラフのみを検索するものとした。

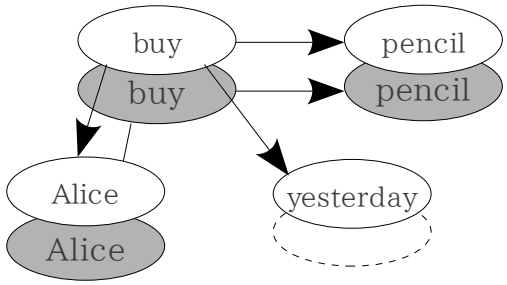

図 3: クエリの部分グラフが検索されない例
白のグラフをクエリ、灰色のグラフをデータベースのグラフとしている。データベースのグラフには yesterday の entity が存在しないため、クエリのグラフの意味に似ているにもかかわらず結果からは除外される

3.2.4 結果のランキング

以上の変更を加えて得られた結果テーブルは、そのままではクエリのグラフに一致するものからほとんど一致しないものまでが無作為に並んでいるだけであり、検索の結果としては用を為さない。これをクエリの意味に近い順にランキングするために、定義ラベル、relation の種類それぞれの条件に類似度を設定する。本研究では上位語は一律 0.2 とするなど仮の値を用いたが、全ての類似度は容易に変更できる。結果テーブルの行は各定義ラベル、各 relation の類似度の和を点数として持ち、結果テーブルはこの点数によってソートされる。これによって、クエリのグラフの内容に似た順に結果をランキングすることができる。

4. 実験

定義ラベル、relation の条件緩和は既存のシステムで発行される SQL 文と大差ないため、検索に大きな差異はないと思われる。しかし部分グラフの検索に対してはほとんど異なる SQL 文を生成するため、検索時間がどれほど異なるか実験を行い比較した。EOLSS を CDL に変換したファイルを mySQL

データベースに格納し、Entity の総数 1,332,746 個、トリプルの数 927,902 個のデータを対象に検索を行った。

```
GET ENTITY WHERE
{?
  <?5:balance(icl>situation)>
  <?1:method(icl>way).@entry.@def>
  <?4:physical(opp>spiritual(aoj>thing))>
  <?D:statistical(aoj>thing)>
  <?U:*>
  [?1 mod ?5]
  [?5 mod^ ?U]
  [?4 aoj^ ?U]
  [?D aoj ?U]
}
```

図 4: 実験した CDQL クエリ

図 4 に示したクエリを

(1)部分グラフを検索しない場合

(2)?U を含んだ部分グラフを検索する場合

の二つの場合でそれぞれ 10 回実行し、その検索時間を比較した。結果は以下のようになった。

(1)8432,26,27,27,26,27,26,27,26,27,26
(2)61206,498,499,500,506,499,501,498,498,500 (単位:ms)

部分グラフは検索できていたが、20 倍ほどの時間が必要になることが判明した。この検索時間の違いは、SQL は結果を求める際に一時テーブルをメモリ上に作成するが、(2)の場合はそれが巨大なため一度 HDD に保存するからであると考えられる。よって、HDD へのアクセス速度をあげる、あるいはメモリを増設しディスクアクセスを回避するなどの方法で速度の改善を期待できる。

なお、最初の一回は双方ともに遅いが、これは HDD からデータをフェッチするためと思われ、二回目以降はメモリにデータがキャッシュされるため速度が改善されている。これもデータベースそのものの チューニングにより改善できると思われる。

5. まとめ

本稿では概念記述言語 CDL で表現されたデータに対する検索を、より意味的に拡張して行う手法を提案した。この手法により、求める情報に対してクエリが多少不適切であった場合も目的の entity を得ることができ、情報を抽出することが可能になる。

課題としては、定義ラベルの拡張の際に品詞情報を使った改良、本稿の実験では仮の値を用いた各類似度のパラメータの調整、実行時間の改善、などが挙げられる。特に実行時間に関してはクエリによって大きく変動するので、発行する SQL 文そのものの改良も検討する必要があると考えられる。

6. 参考文献

[石塚 2009] 石塚満,内田裕士,横井俊夫:自然言語テキスト意味概念の共通的記述による次世代 Web 基盤, 知能と情報(日本知能情報ファジィ学会誌), Vol.21, No.4, pp519-526 (2009)

[堀内 2012] 堀内暢之, 高山智史, 石塚満: 概念記述言語 CDL データの意味的検索法, 情報処理学会全国大会講演論文集, Vol.74, No.1, pp687-688 (2012)

[ISeC 2007] ISeC: CDL.core 仕様書 第 1 版, http://www.instsec.org/CDLcoreSpecV1.pdf