概念グラフのマッチングによる効率的な意味検索手法の提案

Semantic retrieval approach by efficient matching of conceptual graphs

高山 智史 Satoshi Takayama
石塚 満 Mitsuru Ishizuka

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

The amount of information generated by human beings is explosively increasing. Hence, the traditional information retrieval by keywords is getting insufficient to find on-demand information from such a flood of data accurately, and new approaches which consider semantics are required. In this paper, we propose a method for semantic text retrieval by efficient matching of conceptual graphs.

1. はじめに

近年，人類によって生成される情報の量は爆発的に増加しており，そのような大量の情報の中から必要な情報を的確に探し出すことが困難になりつつある．例えば，従来のキーワードによるテキスト検索では，要求に適合する情報を十分に絞ることができず，人手による検索結果の取捨選択に多くの時間が費やされている．このような問題を解決するには，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，意味を考慮した検索の実現に向けて，概念グラフ化されたテキストに対する効率的な検索手法を提案する．概念グラフを用いることで，意味を考慮した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概念グラフを扱うことで，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 章で概念グラフについて説明し，3 章で提案手法について述べ，4 章で関連研究との比較を行い，5 章でまとめと今後について言及する．

2. 概念グラフ

本研究ではテキストを Conceptual Graphs[Sowa 01]をベースとした概念グラフとして扱う．概念グラフは概念を表すノードと概念間の関係を表すエッジからなるグラフ構造ベースの知識表現手法である．例文を概念グラフで表した例を図 1 に示す．

例文: John bought a laptop yesterday.

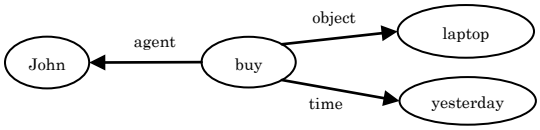

図 1: 概念グラフの例

テキストを概念グラフとして扱うことで，従来のキーワード指定による検索システムでは不可能であった語と語の間の意味的役割関係を考慮した高度な検索条件指定が可能になる．

3. 提案手法

3.1 要件

本研究で実現すべき要件を以下に挙げる．

- 概念グラフ化されたテキスト集合に対し，効率的なグラフマッチングによって意味を考慮した検索を行う．
- 概念グラフのノードに付随する語に対し，その語の同義語，上位語，下位語などを考慮したクエリ拡張を行う．

3.2 定義

本提案で扱う対象，用語について以下のように定義する．(1)概念の種類はあらかじめ定義されており，概念は固有の ID（概念 ID）を持つ．複数の意味を持つ語は意味ごとに概念 ID を持ち，語の意味の曖昧性は解消されているものとする．(2)関係の種類はあらかじめ定義されており，関係は固有の ID(関係 ID)を持つ．(3)グラフの構造を記述するために，便宜的に各ノード，エッジにユニークな ID を付与する（ノード ID，エッジ ID）．(4)検索対象となるテキスト集合から作成された概念グラフをデータグラフ，クエリとして入力される概念グラフをクエリグラフとする．(5)各データグラフに対しグラフ ID を付与する．

以上の定義に従って表現した概念グラフの例を図 2 に示す．（以降，このグラフの ID を G1 とする．）

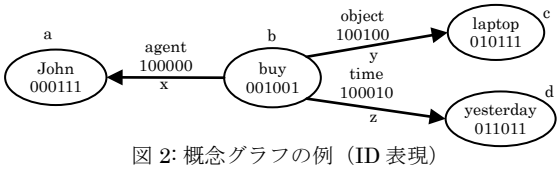

図 2: 概念グラフの例（ID 表現）

3.3 概念 ID

概念に ID を割当てるにあたって，WordNet のような上位語，下位語の情報を持つ木構造のシソーラスを用いる．木構造の各ノードには IP アドレスのようなサブツリーのネットワークアドレスに相当するアドレスと，ネットワークアドレス部を識別するためのビットマスクを付与する．また，補助的な情報として木構造の各階層に階層の深さを表す番号を付与する．この各ノードに割当てられたアドレスをノードに対応する概念の概念 ID とする．概念 ID の割当ての例を図 3 に示す．

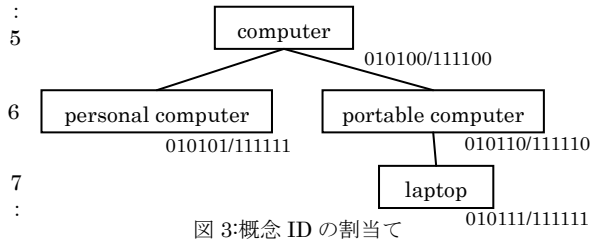

図 3:概念 ID の割当て

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

3.4 グラフ検索

概念グラフ化されたテキストを検索するということは，データグラフの中からクエリグラフを含むものを探すサブグラフマッチングを行うことになる．グラフマッチングは情報科学における基本的な問題のひとつであり多くの研究がなされている．特にサブグラフマッチングは一般に計算量が非常に膨大になるため，マッチングを行う前にフィルタリングを行い，いかにマッチング候補を減らすかが重要になる．本提案では高速なグラフマッチング手法のひとつである GrepVS [Recupero 09]を拡張した手法を用いる．

3.5 インデックス作成

GrepVS では事前にデータグラフに対して fingerprint 情報と構造情報をインデックス化する．

fingerprint 情報はグラフの各ノードを起点とした長さ L 以下の全パスの出現頻度のリストである．L=1 の場合のグラフ G1 の fingerprint 情報を図 4 に示す．

			G1	G2	G3
001001	100000	000111	1	0	･･･
001001	100100	010111	1	1	･･･
001001	100010	011011	1	1	･･･
001011	100000	000110	0	1	･･･
････	････	････	･･･	･･･	･･･

図 4: fingerprint 情報

構造情報はグラフの各ノードを起点とした長さ L 以下の全パスに対する三つ組(ノード ID,エッジ ID,ノード ID)のリストである．L=1 の場合のグラフ G1 の構造情報を図 5 に示す．

			G1	G2	G3
001001	100000	000111	bxa	-	･･･
001001	100100	010111	byc	prq	･･･
001001	100010	011011	bzd	sut	･･･
001011	100000	000110	-	qvs	･･･

図 5: 構造情報

3.6 グラフフィルタリング

与えられたクエリグラフに対し，GrepVS ではデータグラフに対して fingerprint によるフィルタリングと構造情報によるフィルタリングを行うことで候補を絞ってからグラフマッチングを行う．ただし，本提案では fingerprint 情報，構造情報を取得する際のキーとして概念 ID を使用することで上位語，下位語などへのクエリ拡張を実現する．クエリ拡張を考慮したフィルタリングの例を図 6～図 9 に示す．クエリグラフ G_q の概念 ID010100 に対し 010100～010111 へ範囲を広げることで下位語へのクエリ拡張を行っている．

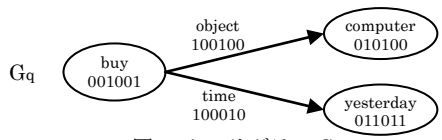

図 6: クエリグラフ G_q

			G1	G2	G3
001001	100000	000111	1	0	･･･
001001	100100	010111	1	1	･･･
001001	100010	011011	1	1	･･･
001011	100000	000110	0	1	･･･
････	････	････	･･･	･･･	･･･

図 7: fingerprint によるフィルタリングでデータグラフを限定

			G1	G2
001001	100100	010111	byc	prq
001001	100010	011011	bzd	sut

図 8: 構造情報によるフィルタリングでマッチング対象となるサブグラフを抽出

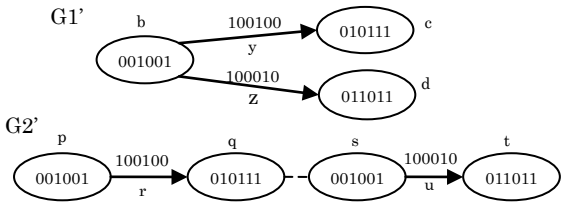

図 9: フィルタリングによって抽出されたマッチング候補

4. 関連研究

[Gómez 00]では，データグラフとクエリグラフを比較して共通部分を抽出し，共通するノードとエッジの数からグラフ間の類似度を計算してランキング を行う手法を提案している．しかし，クエリ拡張や具体的なグラフマッチング手法に関しては言及されていない．

[Zhong 02]ではグラフマッチングの際に WordNet を使ったクエリ拡張を行っている．ただし，グラフマッチングを簡略化するためにあらかじめエントリノードを指定する必要があり，指定できるクエリに制限 がある．

TSUBAKI[Shinzato 08]は日本語の Web ページを対象とした検索エンジン基盤であり，自立語間の係り受けや同義語をインデックス化することで，係り受け関係，クエリ拡張を考慮した検索が可能になっている．しかし，意味的役割関係を考慮していない，クエリ拡張用の同義語をあらかじめインデックスに含めておく必要がある，という点が本提案のモデルとは異なる．

5. おわりに

本稿では，概念グラフの効率的なマッチングによる意味を考慮したテキスト検索方法について提案した．今後は提案手法のパフォーマンス評価を行うとともに，検索結果のランキング法についても検討していく．

参考文献

[Sowa 01] John F. Sowa: Conceptual Graphs , http://www.jfsowa.com/cg/, 2001.

[Recupero 09] Diego Reforgiato Recupero: GrepVS – a Combined Approach for Graph Matching, Journal of Pattern Recognition Research , 2009.

[Gómez 00] Manuel Montes-y-Gómez, Aurelio López-López, and Alexander Gelbukh: Information Retrieval with Conceptual Graph Matching, Proc. of DEXA-2000, 2000.

[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.

[Shinzato 08] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto and Sadao Kurohashi: TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology, Proc. of IJCNLP-08, 2008.