概念グラフのマッチングによる自然言語テキストの意味検索システムの開発

高山 智史† 石塚 満† 内田 裕士‡
†東京大学大学院情報理工学系研究科 ‡UNDL財団

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足であると考える．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の実装として概念グラフのグラフマッチングによるテキスト検索システムを提案する．概念グラフを用いることで，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章で検索システムの概要について説明し，3章で関連研究について述べ，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 概念グラフ

我々のシステムでは自然言語テキストを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による知識表現法である．例文を概念グラフで表した例を図1に示す．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．

例文: John bought a computer yesterday.

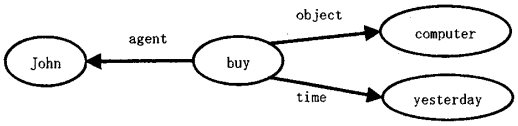

図1: 概念グラフの例

2.2 グラフマッチング

概念グラフ化されたテキストを検索するということは，グラフ集合の中からクエリとなるグラフを部分的に含むものを探す「サブグラフマッチング」を行うことになる．（以降，クエリとなる概念グラフをクエリグラフと呼ぶ．）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的に計算量が膨大になるため，マッチングを行う前にいかに探索空間を減らすかが重要になる．本システムでは，高速なグラフマッチングアルゴリズムのひとつである GrepVS [Recupero 09]を応用しクエリ拡張に対応した方法を用いる．

2.3 クエリ拡張

本システムではクエリグラフの各ノードが表す概念に対して，木構造のシソーラスから上位概念・下位概念を抽出しクエリに追加することでクエリ拡張を行う．木構造のシソーラスの例を図2に示す．

2.4 クエリグラフのノード削除

従来の検索エンジンにおいて，クエリに完全一致する検索結果が得られなかった場合にユーザが検索条件を緩めるのは自然な行動である．

†Satoshi Takayama, Mitsuru Ishizuka (Graduate School of Information Science and Technology, The University of Tokyo)
‡Hiroshi Uchida (The UNDL Foundation)

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.

本システムではこのような要求に対応するために，クエリグラフから末端のノードを徐々に削除して再検索することで十分な量の検索結果が得られるようにしている．

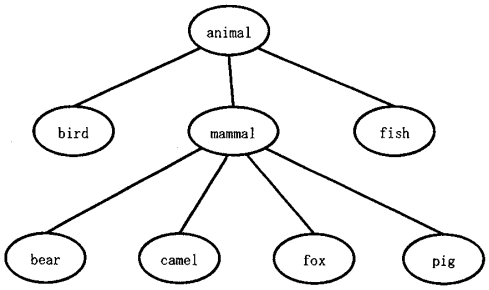

図2: シソーラスの例

2.5 検索結果候補のスコアリング

本システムでは検索結果候補のスコアを以下の3種類の要素から算出する．

(1)TF・IDF[Salton 88]によるスコア

(2)概念の意味的類似性によるスコア

(3)グラフ構造の類似性によるスコア

(1)は従来の検索システムでも用いられる，語の重要性に関するスコアである．(2)はクエリ拡張を行った場合に元の概念と拡張した概念とがどれだけ近いかによって算出されるスコアであり，シソーラスの木構造上での距離が小さいほど大きな値になるように設定している[Zhong 02]．(3)はクエリグラフのノード削除を行った場合に，元のクエリグラフとノード削除を行ったクエリグラフとの間の構造的類似性によって算出されるスコアであり，削除ノード数が少ないほど大きな値になるように設定している．

以上の3つのスコアを合計することで検索結果候補のスコアリングを行い，上位N件を検索結果として出力する．

3. 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，語と語の間の意味的役割関係を考慮していないという点が本システムとは異なる．

4. おわりに

本稿では概念グラフのマッチングによる意味検索システムについて述べた．今後は，クエリグラフのノード削除における削除対象ノードの優先度決定や検索結果候補のスコアリングの重み調整などの課題に取り組んでいく予定である．

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.

[Salton 88] Gerard Salton, Christopher Buckley: Term-weighting approaches in automatic text retrieval, Information Process Management, vol.24, No.5, pp. 513-523, 1988.

[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.

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.