7M-03

# テキストを含む構造化データに対する知識ベースを用いたOLAP

中野茉里香† 阿曽太郎†† 天笠俊之††† 北川博之†††
筑波大学情報学群情報科学類† 筑波大学システム情報工学研究科††
筑波大学計算科学研究センター†††

## 1．はじめに

近年，多くの分野においてテキストを含む構造化データが増加している．一方，データベースに蓄積された大量のデータから多次元的な集計・分析を行う手法としては，OLAP(Online Analytical Processing)が存在する．しかし，テキストに対して OLAP を行なっている従来の研究では語義の曖昧性や表記揺れなどの問題に対応できない上，分析を行うには分野ごとに新たにデータを準備しなければならないといった課題があった．

ところで，知識をデータとして利用可能にした知識ベースが，例えば DBPedia や Wikidata などで，あらゆる分野についてのデータが体系的に記述されており，これを活用した様々な研究が進められている．

そこで，本研究では知識ベースに対するエンティティリンキングを行い，テキストに含まれるエンティティを捉えることで，テキストに対する OLAP を可能にする手法を提案する．抽出したエンティティについて知識ベースに含まれる概念階層などの付加的な情報を用いることで，ユーザが事前に単語の辞書や階層構造の定義を行う必要なしに分析を行うことが可能となる．

## 2．関連研究

TextCube[1]は，あらかじめ定義しておいた単語の辞書，また，単語の意味的な階層構造を用いることでテキストに対する OLAP を実現している．しかし，テキスト中に含まれる単語を扱うため，語義の曖昧性などの問題に対処することができなかった．また，分析したい分野ごとに単語の階層構造を定義しなければならない上，扱うことのできる語彙数に限りがあった．

TopicCube[2]は，テキストから抽出したトピックを新たな次元とすることでテキストに対する OLAP を実現している．しかし，テキストの内容を包括的に捉えたトピックを対象としているため，テキストに含まれる個別の単語について考慮した，より詳細な分析を行うことはできない．また，トピックの意味的な階層構造を事前に定義する必要があるといった問題点があった．

## 3．提案手法

本研究では，テキストを含む構造化データに対して，テキストに含まれるエンティティに着目し，それを知識ベースと関連づけることで，エンティティに着目した OLAP 分析を可能にする手法を提案する．エンティティと知識ベースとの関連づけには，既存のエンティティリンキングの手法を用いる．

エンティティリンキングとは，テキスト中の単語を知識ベース内のエントリと結びつける技術である．与えられたテキスト中の何らかのエンティティを参照する記述であるメンションを検出し，知識ベース内の対応するエンティティと結びつけることができる．エンティティリンキングではテキストからエンティティ名の抽出を行うだけでなく，テキストの文脈を考慮してメンションを知識ベースに結びつけることで単語の表記揺れや語義の曖昧性の解消を行うことができる．なお本研究では，エンティティリンキングの手法は既存のものを用いることとし，その詳細には立ち入らない．

また，テキストから抽出したエンティティについて，知識ベースがもつ概念階層や知識ベースに記述されたエンティティに対する付加的な情報を用いることで，構造化データだけでは不可

**OLAP on Structured Data containing Text using Knowledge Bases**
Marika NAKANO †, Tarou ASO††, Toshiyuki AMAGASA†††, Hiroyuki KITAGAWA†††
† College of Information Science, University of Tsukuba
†† Graduate School of Systems and Information Engineering, University of Tsukuba
††† Center for Computational Sciences, University of Tsukuba

能な OLAP 分析が可能となる．すなわち，テキストから抽出したエンティティの次元に対してもロールアップのような集約の操作が可能となる．

## 4．システムアーキテクチャ

図 1 に本研究で提案するシステムアーキテクチャの概要を示す．入力としては，分析対象となるデータと分析に用いる知識ベースが与えられる．まず，入力の分析対象のデータを格納する．また，選択された知識ベースからエンティティ間の関係についての情報を抽出したエンティティサブグラフ・グラフを生成しておく．次に，エンティティリンカーが分析対象のデータに対してエンティティサブグラフ・グラフを用いてエンティティリンキングを行う．この結果得られた，各レコードに含まれるエンティティとそれに対応するメンションの情報を格納しておく．入力データとエンティティリンキングの結果を用いることで，データの集約を行う．

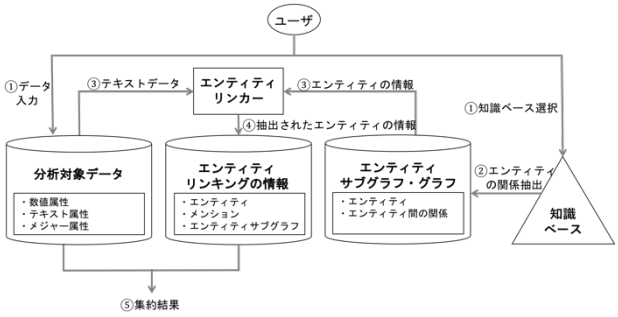


図 1:システムアーキテクチャの概要

## 5. 実験

提案手法について，前処理にかかるコストと精度の評価を行う．比較対象として，従来手法のテキストから単語を抽出し事前に定義しておいた辞書を用いることでテキストの OLAP を行う手法を用いる．

データセットとしては，NHN Japan 株式会社が運営する「livedoor ニュース」のうちクリエイティブコモンズライセンスが適応されるニュース記事を収集したコーパスである，livedoor ニュースコーパス[3]を用いる．このデータセットのうち，スポーツに関するニュース記事 50 件を分析することを想定する．

比較手法における辞書は，対象となるデータに関係するデータとして Wikipedia における「日本のスポーツ選手」カテゴリ下の記事を MediaWiki API を用いて全て収集することで生成する．比較手法の辞書のデータと提案手法のエンティティリンキングによって得られたデータについて比較した結果を表 1 に示す．比較手法がスポーツ選手についての全エントリの情報を抽出しなければならないのに対し，提案手法は抽出されたエンティティの情報のみ抽出するためストレージコストが抑えられる上，実行時間が短くなっている．

表 1:実験結果

| | エントリ数 | データサイズ (byte) | 実行時間(s) |
|---|---|---|---|
| 比較手法 | 55,841 | 11,405,766 | 1680 |
| 提案手法 | 50 | 6,491 | 543 |

また，抽出された単語の分類精度について，比較手法が 44%であるのに対し，提案手法は 72%となっている．

## 6．まとめ

本研究では知識ベースに対するエンティティリンキングによってテキストに含まれるエンティティを捉えることで，より正確なテキストの OLAP を行う手法を提案した．また，テキストから抽出した実体について知識ベースに含まれる概念階層を用いることで，ユーザが事前に単語や単語の階層構造の定義を行う必要なしに分析を行うことが可能となった．評価実験により，提案手法が従来手法よりもコストをかけることなく，高い精度での分析が可能となることが示された．

## 謝辞

本研究では，NHN Japan 株式会社様から提供を受けた「livedoor ニュースコーパス」を利用しました． ここに記して謝意を表します．

## 参考文献

[1] C. X. Lin, B. Ding, J. Han, F. Zhu and B. Zhao, "Text Cube: Computing IR Measures for Multidimensional Text Database Analysis," 2008 Eighth IEEE International Conference on Data Mining, Pisa, 2008, pp. 905-910.
[2] Duo Zhang, Chengxiang Zhai, and Jiawei Han, "Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases," Proceedings of the 2009 SIAM International Conference on Data Mining. 2009, 1124-1135.
[3] livedoor ニュースコーパス.
http://www.rondhuit.com/download.html#ldcc