ISS-SP-393

LOD を用いた非構造データのカテゴライジング手法の提案

槇 俊孝† 若原 俊彦†

† 福岡工業大学大学院工学研究科情報通信工学専攻

1. はじめに

近年，インターネットやネットワーク技術の飛躍的な向上により人々の価値観が多様化し，インターネットはリンク化されているが非構造的なデータが集約したビッグデータで構成されるようになった．新しい技術や価値の創造のためにビッグデータの解析は有効であるが，構造化されていないデータの価値は低いので有効化を図るためには構造化・体系化が重要である．

本研究では，Linked Open Data(LOD)として Wikipedia を活用し，非構造データをカテゴライジングして構造化する手法を提案する．

2. 関連研究

2.1. オープンデータと LOD

オープンデータの普及条件は，ヘテロジニアスなデータを統一的な手段でアクセスでき，そのデータが共通のルールを持つことである．これに対して文書の公開・共有で成功を納めた LOD が注目されている[1]．

2.2. 階層的オートタギング技術

西田氏らは，カテゴリや主題，キーワードを文書に自動的に付与する階層的オートタギング技術を提案し，文書構造を考慮したキーワードタグ抽出法と主題語の抽出精度を評価している[2]．

3. 提案手法

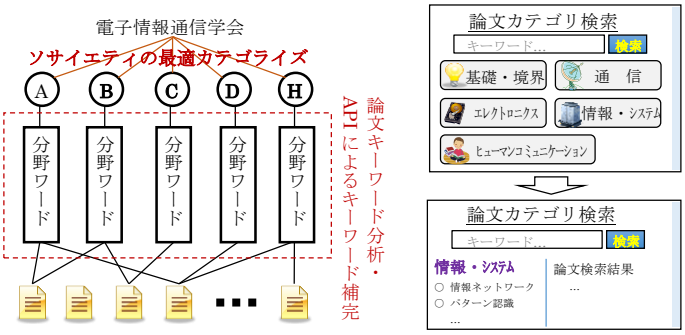
本研究では，Wikipedia リンク API を開発し，論文データを対象としたカテゴライジング手法を提案する．

3.1. Wikipedia リンク API

Wikipedia コンテンツは，公式ページで配布されており自由に利用できる．本研究では，Wikipedia コンテンツをサーバに取り込み，PHP 言語で Wikipedia リンク API を開発した．Wikipedia リンク API は，関連キーワード検索やカテゴリ検索，上位概念キーワード検索，キーワード揺らぎ訂正，日英・英日辞書の機能を実装しており，関連キーワード検索においては，約 8,700 万件のレコードから検索している．図 1 は，Wikipedia リンク API を利用したカテゴリ検索の出力例であり，XML 形式としている．

```
<query>
    <binding name="id">791793</binding>
    <binding name="PoS">一般</binding>
    <binding name="keyword" lang="ja">電子情報通信学会</binding>
    <binding name="keyword" lang="en">
        Institute of Electronics, Information and Communication Engineers
    </binding>
</query>
<category>
    <keyword no="1/6">
        <binding name="id">NULL</binding>
        <binding name="PoS">一般</binding>
        <binding name="keyword">一般社団法人 (学術団体)</binding>
    </keyword>
    <keyword no="2/6">
        <binding name="id">2013743</binding>
        <binding name="PoS">一般</binding>
        <binding name="keyword">日本学術会議協力学術研究団体</binding>
    </keyword>
```

図 1　Wikipedia リンク API によるカテゴリ検索例



(a) カテゴリ階層構造　　(b) カテゴリ検索

図 2　論文データを対象としたカテゴライジング処理

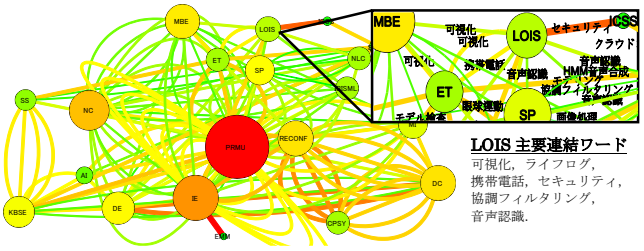


図 3　D ソサイエティにおける LOIS 研究会の関係ネットワーク

3.2. カテゴライジング処理

本研究では，I-Scover[3]に登録されている論文データ約 16 万件 を対象とする．I-Scover は，論文検索において様々なメタデータが示され横断検索ができる非常に有効なシステムであるが，カテゴリ検索の機能が実現されていない．図 2(a)のように分野ワードをカテゴライジングして階層化し，図 2(b)のような検索ができることで，論文検索の効率化を図ることができると考える． これにより 1 件毎の論文の価値が向上し，各論文の比較検討が可能になり，さらに研究専門分野の関係可視化やテクノロジー融合による発想支援が期待できる．図 3 は，論文分析により得られた D ソサイエティにおける LOIS 研究会[4]の分野ワードによる関係ネットワークである．この図より，研究会や専門分野がお互いに密接に関係し，分野が広がっていることが分かる．

4. まとめ

本研究では，Wikipedia リンク API を利用した論文データのカテゴライジング手法を提案し，研究会の関係ネットワーク分析により論文データを分類できることを確認した．今後の課題は，全ての研究会を対象にしてカテゴライジング処理の有効性を評価しカテゴリ検索を実装することである．

参考文献

[1] 大向一輝“オープンデータ活用：1.オープンデータと Linked Open Data”情報処理 54(12), 1204–1210, 2013

[2] 西田京介，星出高秀，藤村考，内山匡“階層的オートタギング技術とその応用”情報処理学会論文誌，データベース 6(1), 29–40, 2013

[3] IEICE Knowledge Discovery http://i-scover.ieice.org/

[4] LOIS 研究会 http://www.ieice.org/iss/ois/jpn/