様式：課程博士用

24年 3月 23日

電子・情報工学専攻	学籍番号	053713
申請者 氏名	小林　暁雄	

指導 教員	増山　繁

論　文　要　旨　(博士)

論文題目	半構造化された文書データを利用したシソーラスの自動拡張及び自動構築に関する研究

(要旨　1,200 字程度)

近年のコンピュータの普及に伴い，自然言語処理によるコンピュータの利便性の向上が望まれている．この際，日々新しく発生する新語を処理するために，それらの新語が収録されたシソーラスの需要も同時に高まっている．本論文では，半構造化された文書から，そのような新語を抽出し，既存のシソーラスを拡張する形でそれらの新語を収録させる手法について解説を行う．本論文では，半構造化された文書として，新語が数多く収録されているウィキペディアと，専門用語が大量に出現する特許文献の電子データを利用する．ウィキペディアは，ウェブ上の大規模な百科事典であり，誰でも記事の作成・編集が可能であるといった，既存の百科事典にはない特徴を備えている．このため，大量のユーザによって，日々新たな記事の追加や，記事内容の検討・修正が行われており，現在，最も多くの記事が収録された百科事典となっている．このような特徴から，ウィキペディアは新語を取得するための最良の文献であり，かつウィキペディアの文書データはウィキソフトウェアによって半構造化されており，データの抽出が比較的容易であるという特性を持っている．このため，本研究では，ウィキペディアから新語を抽出し，既存のシソーラスの意味体系下に分類する形で，シソーラスの拡張を行う．この際に，ウィキペディア記事名のみを抽出するのではなく，記事を分類するウィキペディアカテゴリも同時に抽出し，既存のシソーラスの意味分類体系の下位の分類に相応しいウィキペディアカテゴリのみを選択し，既存のシソーラスの意味分類体系の下位に配置した．こうすることで，誤りが少なく，かつ，大規模な語彙を収録させることが出来ることを確認した．また，この手法を発展させ，ウィキペディアカテゴリ階層から，既存のシソーラスの意味体系の下位の体系として相応しい部分階層のみを抽出し，既存のシソーラスの意味体系下に配置する手法についても提案している．一方，近年の知財訴訟の増加などに伴い，産業界の知財戦略の重要性が高まっており，それに伴い，特許の出願動向の調査における特許文献の解析の需要が高まっている．本論文では，特許出願動向調査における可視化技術の一つである，技術効果型パテントマップに着目した．技術効果型パテントマップは，特定分野における技術とその効果において，どのような特許がどのような技術と効果を持っており，かつ，それに関連・類似した特許の件数などについて，二次元空間上にその分布を表示することで可視化したグラフである．この際，このパテントマップの技術項目を設定するためには，その分野の特許文献に出現する技術用語をまとめ上げ，体系化する必要がある．そこで，本研究では，特許文献に出現する技術用語を単語列とみなして系列マッチングを行い，類義・同義・上位下位関係でまとめたシソーラスの自動構築手法を考案した．