氏　　名	松岡有希
学位（専攻分野）	博士（情報学）
学位記番号	総研大甲第1154号
学位授与の日付	平成２０年３月１９日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第６条第１項該当
学位論文題目	意味的メタデータ生成のための協調型アノテーションに関する研究
論文審査委員	主　査　教授　武田　英明 准教授　市瀬　龍太郎 准教授　相原　健郎 准教授　北本　朝展 准教授　松尾　豊(東京大学)

論文内容の要旨

近年の WWW では，ユーザ参加型のサービスが多く提供されており，多種多様な情報が大量に存在する．大量の情報は，反面ユーザがほしい情報を探すときの妨げになっている場合がある．これを解決する一つの方法として，Web コンテンツにメタデータを付加することが挙げられる．メタデータは“data about data” と定義され，情報の管理，統合，検索などの用途で使用されてきた．本研究では，メタデータを情報検索の目的で使用するために，Web コンテンツの内容を代表する意味的メタデータを生成することを目標とする．

意味的メタデータには，Web コンテンツの主題を表す語や特徴語，コンテンツの内容と関連する語が記述されることが望ましい．意味的なメタデータを生成の二つの課題（1）誰がメタデータを生成するのか，（2）どのようにメタデータを生成するのか，を次のアプローチで解決を図る．（1）Web コンテンツの著者がメタデータを生成する場合，著者の意図が反映されてしまい，読者にとって有益な情報が提供されるとは限らない．そこで本研究では，Web コンテンツの複数の読者によってメタデータを生成することを提案する．複数ユーザによる集合知を活用することで，質の高いメタデータの生成が期待できる．（2）Web コンテンツの読者が負担を感じることなく，メタデータを作成してもらえるようなアーキテクチャが必要である．そこで本研究では，メタデータ生成のためにアノテーションを用いることを提案する．ユーザが読書をする際の自然な行為をメタデータ生成に用いることにより，ユーザのメタデータ生成における負担を軽減することが期待できる．

なお，本稿では人工知能学会全国大会で運用された大会支援システムの一機能として提供したアノテーションシステムの実装・運用によって得られたデータを基に分析を行った．

まず初めに，ユーザが Web コンテンツ内で下線を付与した箇所にはどのような特徴があるのかについて調査した．イロノミーは，三色ボールペン読書法に基づいて下線を付与できるシステムである．分析の結果，全ユーザで見ると，色にかかわらず tfidf 値の高い語，すなわち特徴語に下線が付与される可能性が高いということがわかった．また，下線が付与された語は Web コンテンツの内容を直接反映した語と判断でき，主題を表す語が含まれることから，複数ユーザによって付与された下線文を集約すると，意味的メタデータの生成に利用できる語が含まれることが見出された．

次に，ユーザがマーキングを付与した語や文字列を他人と共有した場合，情報探索に役立つのかについて調べた．ページ間類似度やマーキングされた文字列内（マーキング文字列）の語を使ったページ推薦を行う合口を運用することにより，分析を行った．その結果，ユーザは学会中において，ページ間類似度によるページ推薦よりも，他のページに付与されているマーキング文字列内の語を使ったページ推薦を選択することが示唆された．

最後に，ユーザが発表を聴講している際に書いたメモと論文内容との関係性について調査した．ユーザが発表聴講時に 2 種類のメモ（個人メモ，質問メモ）を書くことができるシステム，memoQ を運用した．ユーザは memoQ を利用して，分析の結果，ユーザがメモを入力するときのコンテキストを利用することによって，メモからコンテンツ内の特徴語やコンテンツに含まれないが内容と関連のある語が獲得できる可能性が見出された．

これらの分析結果より，複数ユーザが付与したアノテーションから意味的メタデータに有用な語を獲得できる可能性があることが分かった．

論文の審査結果の要旨

本論文は，情報探索を向上させるための意味的メタデータを生成する方法について探求したもののである．意味的メタデータとはコンテンツの内容を代表するメタデータのことであり，このようなメタデータを用意することで，情報の検索や理解，再利用が容易になる．既存のメタデータ生成手段を分析した結果，意味的メタデータ生成の手法として，協調型アノテーションによる方法を基本的な枠組みとして提案している．これは，現在ソーシャルタギングとして普及している方法であり，（1）集合知による洗練効果，（2）参加の枠組みによる自然な情報提供を狙ったもので，これによりユーザに過大な負担をかけずに良質な意味的メタデータが生成できるとしている．ただし，既存のソーシャルタギングシステムでは，意味的メタデータでないものを含んでしまう点に問題がある．そこで，本研究ではソーシャルタギングシステムを基本とし，その方法を改良することで意味的メタデータの獲得の可能性を論じている．

方法は多く二つに分かれる．1つはコンテンツに直接含まれる語から意味的メタデータを生成する方法である．本研究ではタギングではなくコンテンツの一部をマーキングすることでメタデータを獲得する方法を調べている．それが第1および第2のシステムに関する研究である．また，コンテンツに直接含まれない語から意味的メタデータを生成する方法に関してはユーザが作るメモ書きを利用する方法を検討している（第3のシステム）．

第1のシステム（「イロノミー」）はマーキングによる情報獲得が適切なメタデータを生成することができるかを分析している．ここでは文書集合全体の中での単語の重要性を図る指標である tfidf による指標と比較している．個々のユーザが生成するマーキング文字列からはtfidf法における分布とは対応しない単語集合が得られるが，ユーザ全体からtfidf法における値の高い単語が文書集合全体よりも頻度高く出現することがわかった．すなわちマーキングはコンテンツ内の重要性の高い単語を抜き出す効果があることがわかった．

第2のシステム（「合口」）では，マーキング文字列が情報探索に有効かどうかを情報推薦の効果を調べることで検証している．結果としては，マーキング文字列を利用した情報推薦は文書全体の単語を利用した推薦や協調的フィルタリングより選好されていることがわかった．

第3のシステム（「MemoQ」）では，コンテンツ内にない情報を獲得する手段としてメモ書きに注目して，メモ書きから意味的メタデータに適切な情報が得られるかを調べている．その結果としては，特に質問メモからは直接コンテンツ内にないが，関連する用語が得られることがわかり，有効な手段であることがわかった．

以上の結果から，協調型アノテーションは意味的メタデータ生成の有効な方法となりうることを示した．

本研究は情報探索が社会的問題となっている現在において有効な解決策となりうる意味的メタデータを如何に獲得可能になるかという野心的なテーマをシステム構築・運用を通じた分析によって探求しており，独創的かつ重要な研究であると認められる．ことに実際の運用を通じて評価を行っている点は大いに評価できる．

以上の点をもって，この論文は博士（情報学）の学位論文として十分な価値があるものとして認められる．