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

松岡有希

博士（情報学）

総合研究大学院大学
複合科学研究科
情報学専攻

平成19年度
(2007)

本論文は総合研究大学院大学複合科学研究科情報学専攻に
博士（情報学）授与の要件として提出した博士論文である。

審査委員：
武田 英明（主査）
相原 健郎
市瀬 龍太郎
北本 朝展
松尾 豊　　東京大学
（主査以外はアルファベット順）

Studies about Collaborative Annotation for Making Semantic Metadata

Yuki Matsuoka

DOCTOR OF
PHILOSOPHY

Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)

March, 2008

A dissertation submitted to
the Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Advisory Committee:
Hideaki Takeda (Chair)
Kenro Aihara
Ryutaro Ichise
Asanobu Kitamoto
Yutaka Matsuo The University of Tokyo
(Alphabetical order of last name except chair)

内容梗概

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．こうした大量の情報は，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，検索などさまざまなま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するために，(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，という2点の課題に対し，それぞれ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) メタデータには，情報探索に役立つ情報が書かれていることが望ましい．また，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．アノテーションとは人が文書を読む際に，紙上に下線やハイライト，メモなどを書き込む行為である．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．また，ユーザがアノテーションによってコンテンツ内で注目した箇所や新たに追加した情報から情報探索に有益な語を獲得できる可能性がある．

本稿では，ユーザがWebコンテンツに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かどうかについて，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．アノテーションシステムは，人工知能学会全国大会（JSAI）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．対象としたWebコンテンツは学会で発表された論文の情報が書かれた発表ページである．

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．JSAI2005で運用されたイロノミーは，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．JSAI2006 において，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．また，ユーザはシステムによって推薦されたページよりも，他人がマーキング文字列から他のページへ張ったリンクを選択することも示唆された．さらに，ユーザが付与したマーキング文字列内の語のうち，tfidf 値の高い語が情報探索に有益かどうかについて調べたところ，ユーザは tfidf 値の高い語とは関係なく，他人が付与したマーキング文字列内の語を利用した情報探索を好むということが分かった．これらの結果より，他人がマーキングによってページ内で注目した語は tfidf 値の低い語でも情報探索に有益であることが見出された．

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．JSAI2007 において，ユーザが発表聴講時に 2 種類の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，発表者に対する質問用のメモ（質問メモ）と，個人用のメモ（個人メモ）の 2 種類を入力できる．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．

Abstract

Recently, there are many kinds of Web pages on WWW, because there are many services based on the architecture of participation. When users look for the information from such many Web pages, users may not get it. To solve the problem, I focus on adding metadata to web contents. Metadata is defined that "data about data", it is used for management of information or unification or search, and so on. In this thesis, I use metadata for information search and propose to generate the semantic metadata based on content. I suppose the semantic metadata includes subjective words or characteristic words, relevant words to words in Web contents. To generate semantic metadata, I focus on the approach for two points of problems; (1)who generates metadata and (2) how generates metadata.

(1) When a creator of the Web content generates metadata, the metadata may not be useful for readers. I propose that readers of the Web content generate metadata. If readers generate metadata with the wisdom of clouds, generated metadata may be high quality.

(2) Semantic metada should be written information to help the information search. I have to propose the architecture for generating metadata without user costs. I propose to use annotation for generation semantic metadata. Annotation includes underlining or highlighting in Web contents and adding memo on Web contents.

I analyzed whether annotation which users give to Web contents are useful for generating semantic metadata. I offered the annotation systems as one ability of the web support system managed by conference of the Japanese Society for Artificial Intelligence (JSAI). I use the presentation pages include information of papers presented in the conferencethe as web contents.

At first, I investigated features about underlined words in Web contents. *Ironomy* operated in JSAI2005 is the system which users can underline words based on a reading method using three color ballpoint pen. I understood underlined words by all users include words with high tfidf value. In addition, I found that underlined words are direct content-based and include some subjective words. As a result, underlined words by all users include useful words for generating semantic metadata.

Next, I investigated whether sharing marking data is useful for information search. *Aikuchi* operated in JSAI2006 is the system which recommend pages based on four algorithms. As a result of analysis, users chose recommended pages using marked strings than using page similarity. In addition, I analyzed whether users prefer words with high

tfidf value for information search. Then I found that users preferred marked words with low tfidf value. Sharing marked words is useful for information search.

Finally I investigated relationship between memos and web contents, the memos were written when users listen to the presentations. In JSAI2007, I managed system, *memoQ* that users could input two kinds of memos at the listening to the presentation. The user can input two kinds of a memo for questions (a question memo) and a memo for the themselves (a personal memo) by using memoQ. As a result of analysis, I found that question memos include words related to the web page without being included in the page, and personal memos include words directly related to the web pages.

I found that annotation which users gave to web content include useful information for information search.

目 次

図目次

表目次

第 1 章
序言

1.1 本研究の背景と目的

Tim O'Reilly らが提唱した Web2.0[O'Reilly 05] の登場に伴い，近年，WWW 上における情報量が増大している．Web2.0 にはさまざまなキーワードが含まれているが，その中の一つに，ユーザ参加アーキテクチャがある．ユーザ参加型アーキテクチャとは，ユーザが一方的に情報を受け取るだけでなく，情報を発信することができるようなアーキテクチャのことである．Web2.0 が提唱される以前は，ユーザが自ら HTML を使って Web コンテンツを作成する必要があったが，Blogger[*1] や WordPress[*2] のような Weblog のホスティングサービスやソフトウェアを代表するように，HTML の知識がなくても簡単に Web コンテンツを作成することができるシステムが設計されるようになった．ユーザは Weblog を利用することによって，HTML のフォームにタイトルや記事を入力するだけで簡単に Web コンテンツを作成することができる．このように，ユーザが簡単にコンテンツを投稿できるシステムが登場したことにより，今まで情報を受け取るだけであった一般ユーザが積極的にコンテンツを作成するようになった．Weblog の一般的な定義は，Permalink と呼ばれる永続的な URI を持ったコンテンツが時系列に表示されるサイトである [武田 04]．コンテンツの内容は，ニュースや政治に関する意見や，製品に関するレビュー，日記など，様々である．その他，文章ではなく，写真を投稿することができる Flickr[*3] や，動画を投稿することができる Youtube[*4] といったシステムがある．これらのシステムにおいても，ユーザが簡単に写真や動画を投稿できるように設計されている．こうしたユーザ参加型アーキテクチャを取り入れたサービスの普及が，多種多様な情報を爆発的に増加させる一因となっている．

[*1] http://www.blogger.com/
[*2] http://wordpress.org/
[*3] http://www.flickr.com/
[*4] http://jp.youtube.com/

これらWWW上にある多種多様な大量の情報は，ユーザがほしい情報を探すときに大いに役立っているが，情報量が多すぎて情報探索に失敗する場合も生じている．WWWにおいてユーザが行う情報探索には，大きく分けてknown-item searchとsubject searchの2種類がある[Matthews 83]．known-item searchは，ユーザが獲得したい情報が的確に決まっているときの情報探索である．たとえば，特定の電化製品の評判について調べるときや，誰かの所属について調べるとき，といった状況における情報探索である．一方のsubject searchは，主題に関する情報を探すときの情報探索である．たとえば，セマンティックWebに関する文献を探したいときや，神保町駅付近にあるレストランを探すとき，といった状況における情報探索である．

現在，WWW上でユーザが情報を探す際の手段としては，検索エンジンを利用する，情報推薦サービスを利用する，等がある．これらの情報探索手段において，known-item searchとsubject searchといった情報探索をする際，さまざまな問題が生じている．

Google[*5]やYahoo![*6]が提供する検索エンジンは，ユーザが検索クエリと呼ばれるキーワードを入力として受け取ると，検索クエリに基づいてユーザに検索結果の一覧を返す．検索エンジンは，WWW上にあるWebコンテンツを定期的に収集し，自動的にWebコンテンツのインデックスを作成する．ユーザが検索クエリを使って検索エンジンに問い合わせると，検索エンジンは事前に作成したコンテンツのインデックスやコンテンツ間のリンク関係などを利用して，ユーザに検索結果の一覧を返す．現在，WWW上でユーザが情報を探すときにもっとも利用するのは検索エンジンである．図1.1は検索エンジンによる情報検索のベン図である．AはWWW上の全てのコンテンツ集合で，Bは検索結果のコンテンツ集合，Cはユーザがほしいコンテンツ集合を示している．多くの検索エンジンは検索クエリがコンテンツに含まれるかどうかによる全文検索をしているため，検索結果にはユーザがほしいコンテンツ以外のものが多く含まれてしまうという欠点がある．例えば，ユーザが意図していたのとは異なる意味の検索クエリが含まれるコンテンツや，意図していた検索クエリが含まれるもののユーザにとって必要でないコンテンツが含まれることがある（図1.1のうち，B\Cの部分）．ユーザが検索エンジンを使って目的の情報を探す際（known-item search），知りたいことが明確に分かっているため，的確な検索クエリを使うことができるが，ユーザがほしいコンテンツとは異なるコンテンツが含まれるため，検索エンジンが返した結果一覧の中から目当てコンテンツを見つけなければならない，という問題がある．一方で，ある主題に関する情報を探す場合（subject search），もしユーザ

[*5] http://www.google.com/
[*6] http://www.yahoo.com/

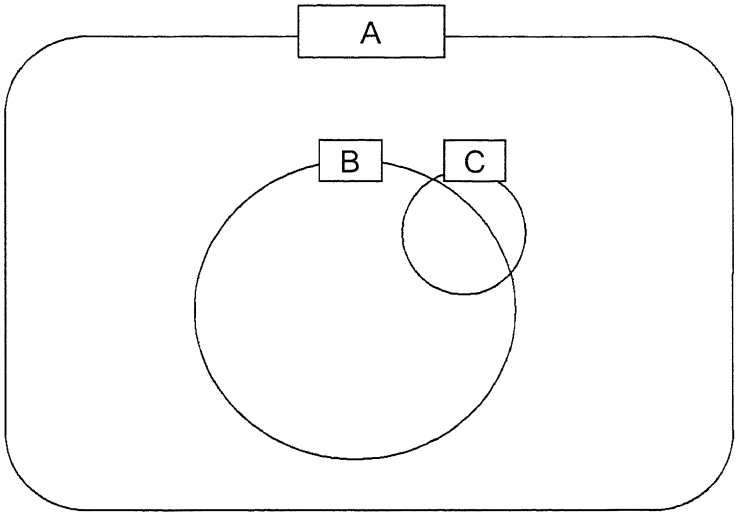

図 1.1: 情報検索におけるベン図　A：WWW 上の全てのコンテンツ集合，B：検索結果のコンテンツ集合，C：ユーザがほしいコンテンツ集合

が主題に関する知識がなければ，検索エンジンによって返されたコンテンツに価値があるかどうかの判断が難しいことがある．図 1.1 のうち，C\B の部分は，全文検索では検索されないコンテンツ集合である．この集合には，検索クエリと関連のある語を含むコンテンツや，検索クエリとは全く異なる意味の語を含むコンテンツがあるものと思われる．

情報推薦サービスとは，ユーザの嗜好に合わせて情報を推薦するサービスである．情報推薦サービスで使われている技術は，コンテンツに基づくフィルタリング（Content-based filtering）と，協調フィルタリング（Collaborative filtering）の 2 種類がある [土方 04]．コンテンツに基づくフィルタリングでは，検索エンジンの検索結果として出力された複数のコンテンツの中からユーザが選択したコンテンツの内容に基づいて他のコンテンツを推薦したり，検索クエリに含まれる語のベクトルとコンテンツ内に含まれる語のベクトルが似ているものを推薦する．協調フィルタリングでは，Web コンテンツに対するユーザの明示的あるいは暗黙的な評価を用いることによってユーザプロファイルを作成し，似た嗜好を持つユーザが評価したことのあるコンテンツを推薦する．ユーザのコンテンツへの評価方法は，ユーザが Web コンテンツに対して数値による評価をしたり，ユーザの Web コンテンツへのアクセス履歴や閲覧時間などを評価に利用したりする．情報推薦サービスで代表的なサービスは，Amazon[*7] である．Amazon ではユーザがある商品のページを見ると，「この商品を買った人はこんな商品も買っています」という表示と共に，他の商品を推

[*7] http://www.amazon.com

薦する．コンテンツに基づくフィルタリングの場合，ユーザが調べたい情報が明確なとき（known-item search）に検索クエリと似たコンテンツが返されることや，関連する情報を探したいときに（subject search）現在見ているコンテンツと似たコンテンツを取得できることは利点であるが，コンテンツ集合があらかじめ分かっていないと計算ができないため，広い WWW 空間には適さない．協調フィルタリングの場合，ユーザが興味を持っているコンテンツと関係のあるコンテンツが推薦されるため，まったく新しい分野のコンテンツを探すことには適さない．

このように，現在の WWW では，情報を探すときに利用できるサービスがいくつかあるが，ユーザが情報を探す際に様々な問題が生じている．解決策の一つとして，Web コンテンツに付与されたメタデータを利用することが挙げられる．もし，Web コンテンツに意味的メタデータが付与されていれば，known-item search と subject search の問題が解決できる．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る．意味的メタデータには，コンテンツに含まれる語だけでなく，コンテンツに含まれない語も記述される．しかし，コンテンツの内容とまったく関係のない語が記述されることは望ましくない．意味的メタデータとして，Web コンテンツの内容を直接反映した語や，Web コンテンツ内の語の意味の定義などが書かれていることが望ましいと考えている．Web コンテンツの内容を直接反映した語とは，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のことである．また，情報価値のある Web コンテンツにだけ，意味的メタデータを生成することが望ましいと考えている．もし，意味的メタデータがあれば，図 1.1 において，B\C の部分に含まれていたユーザが意図していたのとは異なる意味の検索クエリが含まれるコンテンツや，ユーザにとって必要でないコンテンツが検索結果に含まれなくなるため，ユーザが意図していた意味の検索クエリが含まれるコンテンツを獲得できるようになる．したがって，known-item search の問題を解決することができる．また，主題語や特徴語から Web コンテンツの要点を知ることができるようになるので，subject search における問題も解決することができる．C\B の部分からは，検索クエリと関連のある語を含むコンテンツを獲得できるようになるだろう．情報推薦サービスにおいては，意味的メタデータに主題語や特徴語が書かれているため，コンテンツのプロファイル作成時の計算が高速になる可能性がある．しかし，現状では情報探索に役立つ意味的メタデータは普及していない．そこで本研究では，ユーザがほしい情報を獲得しやすくなるような，意味的メタデータを生成・流通させることを目的とする．

1.2 メタデータについて

メタデータは一般的に *data about data* と定義されている．簡単な定義であるが故に，メタデータはさまざまな目的で使われ，メタデータの付与対象物もさまざまである．したがって，人によってメタデータの意味合いが異なる．そこで，これまで生成されてきたメタデータの役割や記述内容について紹介するとともに，本研究で目標としている意味的メタデータについて明らかにする．

メタデータは情報資源について言及したものであり，下記に示した 3 種類の特徴を持ったメタデータがある [Baca 98]．

Contet

情報資源に含まれる要素や，情報資源について述べたメタデータ

Context

誰が，何が，なぜ，どこで，どのように情報資源と関連があるのかを示すメタデータ

Structure

個々の情報資源内あるいは情報資源間の連携の形式に関するメタデータ

メタデータは古くから図書館や博物館において資料の記録保持や管理のために使われていた．したがって，図書館や博物館におけるメタデータは，*Context* に焦点を当てていた．メタデータの記述標準化についてはアメリカの図書館業界が早くから行っており，一番最初の標準化は 1984 年に公表された，MARC Archival and Manuscript Control (AMC) である．MARC-AMC は，インデックスや概要，目録などの構造や規格に関する標準化であった．しかし，MARC-AMC では，資料の階層性を表す記述に限界があるという問題があった．後に，コンピュータ環境の発展によって情報共有の必要性が高まり，1994 年に Encoded Archival Description (EAD) が策定された．EAD は SGML/XML DTD (Document Type Definition) として開発されているため，より柔軟な階層記述ができるようになった．これにより，*Structure* の機能が強化された．*Content* に関する代表的な標準規格は，Dublin Core である．Dublin Core は，1995 年ころから Dublin Core Metadata Initiative (DCMI) によって策定されたものであり，情報資源の発見を目的としている．Dublin Core は，Simple Dublin Core と呼ばれる 15 の要素からなる（表 1.1[日本 04]）．

表 1.1: Dublin Core の 15 の基本要素

エレメント名	日本語の表示名	定義および説明
Title	タイトル	情報資源に与えられた名称
Creator	作成者	情報資源の内容の作成に主たる責任を持つ実体
Subject	キーワード	情報資源の内容のトピック
Description	内容記述	情報資源の内容の説明・記述
Publisher	公開者	情報資源を公開することに対して責任を持つ実体
Contributor	寄与者	情報資源の内容に何らかの寄与，貢献をした実体
Date	日付	情報資源のライフサイクルにおける何らかの事象の日付
Type	資源タイプ	情報資源の内容の性質またはジャンル
Format	記録形式	情報資源の物理的形態ないしディジタル形式での表現形式
Identifier	資源識別子	当該情報資源が作り出される源になった情報資源への参照
Source	出処	当該情報資源の知的内容を表す言語
Language	言語	当該情報資源の知的内容を表す言語
Relation	関係	関連情報資源への参照
Coverage	時空間範囲	情報資源の内容が表す範囲または領域
Rights	権利管理	情報資源に含まれる，またはかかわる権利に関する情報

これらをまとめると，最初は個々の図書館や博物館における資料の管理のための *Context* に特化したメタデータが必要とされ，ネットワークの普及によって情報統合の要求が高まったことから *Structure* に特化したメタデータが登場し，さらに高度な情報検索のために *Content* に特化したメタデータが必要とされてきたことが分かる．メタデータの付与対象となるのは，本や美術資料などがデジタル化された資料，Web コンテンツなど，情報的価値のあるものなら何でもかまわない．本研究では，WWW 上の情報資源を対象としている．すなわち，HTML で記述された Web ページや論文などの PDF ファイルや写真，動画といった Web コンテンツを想定している．また本研究では，多種多様な大量の情報の中から有益な情報を探すためにメタデータを利用しようとしている．したがって，3 種類のメタデータのうち，*Content* に特化したメタデータに焦点をあてる．

メタデータに記述される内容は，[Kashyap 97] が具体的な例を用いて下記のように分類している．

Content Independent Metadata

コンテンツの内容と直接関係がないもの　(ex. コンテンツが生成された場所や編

集日）

Content Dependent Metadata

コンテンツの内容と関連があるもの （ex. コンテンツのサイズや語数）

Direct Content-based Metadata

コンテンツの内容を直接反映しているもの （ex. コンテンツの記述に基づいて作成されたテキストのインデックス）

Content-descriptive Metadata

コンテンツの内容を直接利用することなく内容に関することが記述されているもの（ex. テキストアノテーション）

これらのメタデータの使用方法としては下記の用途がある [Stuckenschmidt 04].

Structuring

メタデータはトピックエリアやキーワードや他の情報との関連性を記述することによって，さまざまな情報を構成するために使うことができる

Maintenance

メタデータは著者や作成日，期限などを書くことによってコンテンツのメンテナンスを助けることができる

Cataloguing

大きい情報レポジトリはコンテンツの概要を持っていることがさらに重要になる

Search

メタデータにトピックエリアやキーワード，サマリーを記述することによって，Web ページを一つずつ検索することなく，Web 上のコンテンツを特定するために使うことができる

Access

フォーマットやエンコーディング，ツールやラッパーへのリンクのようなコンテンツの技術プロパティに関連するメタデータは利用できるコンテンツを処理しやすくなる

Interpretation

使用されている専門用語や作られた条件，コンテンツを解釈するのに必要とする知識は，コンテンツを本当に理解させるために人間とシステムの両方にとって必要である

また，メタデータの用途と前述した分類の関係を表 1.2 に示す.

表 1.2: メタデータの用途と分類の関係

用途	分類
Structuring	Direct Content-based Metadata, Content-descriptive Metadata
Maintenance	Content Independent Metadata
Cataloguing	Direct Content-based Metadata, Content-descriptive Metadata
Search	Direct Content-based Metadata, Content-descriptive Metadata
Access	Content Independent Metadata, Content Dependent Metadata
Interpretation	Content-descriptive Metadata

メタデータを Structuring や Cataloguing，Search，Interpretation のために使用する場合，コンテンツ内のテキストから作成したインデックス（Direct Content-based Metadata）や，アノテーションによって付与されたコンテンツの内容と関連のあるキーワード（Content-descriptive Metadata）などが必要である．一方で Maintenance や Access のために使用する場合は，コンテンツの内容とは直接関係のない著者や作成日（Content Independent Metadata），コンテンツのサイズ（Content Dependent Metadata）などが書かれたメタデータが必要である．両者のうち，多種多様な情報の中からユーザほしい情報を獲得するのに役立つのは，前者の Direct Content-based Metadata や Content-descriptive Metadata が該当する．Structuring の要件を満たしたメタデータがあれば，ユーザは一つの Web コンテンツから関連する他の Web コンテンツを獲得することができるようになるため，情報探索の手間を省くことができる．Cataloguing の要件を満たしたメタデータは，カテゴリ分類が記述されたメタデータなので，ユーザはカテゴリにアクセスするだけで簡単に大量の Web コンテンツを獲得することができるようになる．Search の要件を満たしたメタデータがあれば，ユーザは Web コンテンツが探しやすくなる．Interpretation の要件を満たしたメタデータは，Web コンテンツに含まれる語や内容に関する説明が記述されたメタデータなので，ユーザが該当コンテンツに関する背景知識を持っていないときに役立つ.

したがって，本研究で目標としているメタデータは，情報探索を目的とした *Content* に特化

したメタデータであり，Direct Content-based Metadata や Content-descriptive Metadata といった，Web コンテンツ内のテキストやアノテーションから得られる語が記述されたメタデータである．記述される内容に関しては，Web コンテンツの主題を表す語や特徴語，内容と関連する語，コンテンツ内の語の意味の定義などが記述されていることが望ましいと考えている．

1.3 メタデータ生成における問題点

現在，WWW 上では大きく分けて 4 種類のメタデータが存在する．これらのメタデータを，生成方法について分類すると図 1.2 のようになる．まず，メタデータを生成する人としてコンテンツの著者と読者がいる．また，メタデータは自動的に生成される場合と，人が手動で生成する場合がある．本節では，4 種類のメタデータを検証することで，1.2 で述べた，本研究で目標としている意味的メタデータを生成する上での問題点について述べる．

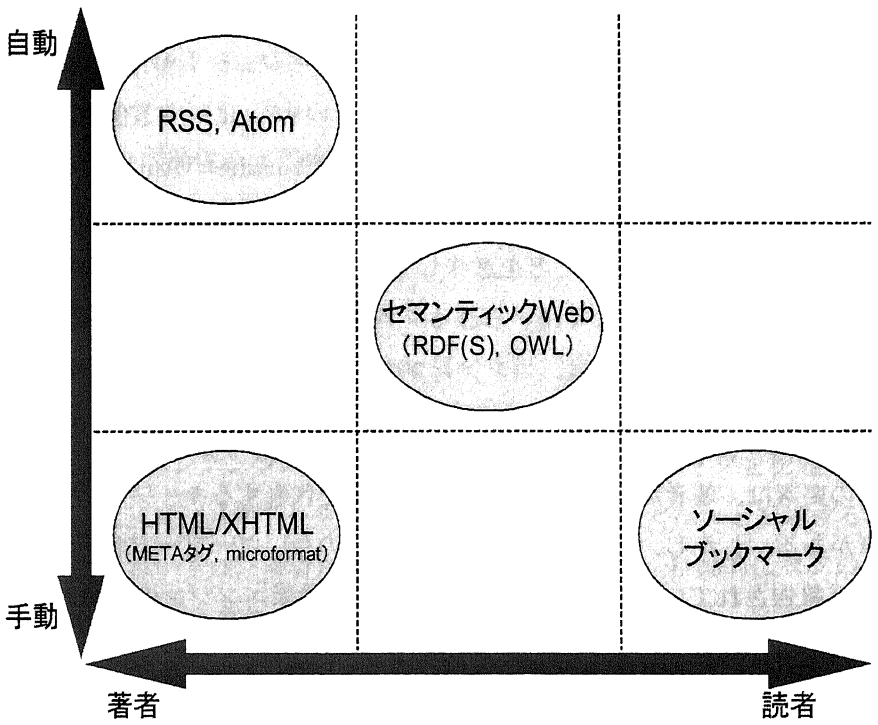

図 1.2: メタデータの生成方法に関する分類

1.3.1 HTML/XHTML におけるメタデータ

HTML[Dave Raggett 99] では，コンテンツの著者がメタデータを記述できるような要素が下記のように用意されている．

address タグ

address タグは bod y 部分に記述することができ，コンテンツの著者の名前や連絡先が記述される．著者がコンテンツの内容の問い合わせ先を提供するために使用される．

title 属性

リンクや画像，あるいは文字列に関する説明を記述するために使用される．コンテンツ内のリンクや画像の上にマウスカーソルをもっていくと，title 属性に記述された内容がツールチップで表示される．

META タグ

META タグはヘッダ部分に記述され，ユーザエージェントが META タグの内容を読み込む．META タグでは，プロパティとプロパティに対する値を記述することができる．例えば，HTML のヘッダ部分に <META name="Author" content="Yuki Matsuoka"> と記述した場合，これはコンテンツの著者（プロパティ）は Yuki Matsuoka（値）であるということを意味している．name 属性は任意のプロパティを記述することができる．よく用いられるプロパティは，description や keywords である．これらのプロパティは，コンテンツの内容を表すキーワードや要約を書くために使用される．

これらの要素は，著者が自由にコンテンツの内容を代表するキーワードを記述できるという利点がある．しかし，コンテンツの内容と関係のないキーワードが記述されることが多いことが報告されている [Lawrence 00]．これは，検索エンジンの結果が上位に表示されるようにするために，著者がわざと META タグにコンテンツの内容と関係のないキーワードを記述するからである．そのため，META タグに記述されたキーワードの優先度を低く設定するようになった検索エンジンもある．したがって，META タグに記述された内容は必ずしもコンテンツの内容が直接反映された情報でないという欠点がある．

他に著者がコンテンツにメタデータを記述する方法としては，microformats[Khare 06] がある．microformats では，HTML または XHTML[Pemberton 02] において，CSS（Cas-

cading Style Sheet）[Meyer 01] で用いられる class 属性に，microformats のために提供されているフォーマットに基づいて，class 属性値と属性値に対する値を記述することによってメタデータを生成する．例えば，連絡先情報を記述するフォーマット hCard[*8] に基づいてメタデータを記述すると，下記のようになる．

```
<div class="vcard">
    <div class="fn">Yuki Matsuoka</div>
    <div class="org">SOKENDAI</div>
    <div class="tel">03-4212-2681</div>
</div>
```

hCard は，vCard[Dawson 98] のオブジェクト/プロパティ名を小文字にしたものを class 属性の値として使用している．fn は表示名，org は所属，tel は電話番号を意味する class 属性値である．全体は class=“vcard” でラップされており，これらのクラスが hCard(vCard) を構成している．このように，著者がコンテンツにフォーマットに基づいたメタデータを生成すると，ソフトウェアエージェントが自動的に情報を抽出することができるようになる．その他のフォーマットは，カレンダー・イベント情報配信用の hCalendar[*9] やハイパーリンクを利用して人間関係を表現する XFN (XHTML Friends Network)[*10] などがある．XFN を利用したソフトウェアエージェントとして，Google が Social Graph API[*11] を提供している．

Web のコンテンツの内容を直接反映させたメタデータを作成するためには，rel-tag[*12] というフォーマットが用意されている．rel-tag はハイパーリンクに rel=“tag” を書き加えることにより，Web コンテンツに対して著者が定義したタグ（キーワード）を関連付けることができる．Web コンテンツ内に著者が「メタデータ」というタグを記述するときは， メタデータ と記述する．一部の Weblog サービスにおいては，著者が記事にタグを付与すると，自動的にタグのハイパーリンクに rel=“tag” が挿入されるようになっている．

このように，HTML や microformats には，著者がコンテンツの内容を反映したインデックスや説明を記述できる要素が用意されている．しかし，著者が Web コンテンツの内容

[*8] http://microformats.org/wiki/hcard
[*9] http://microformats.org/wiki/hcalendar
[*10] http://www.gmpg.org/xfn/
[*11] http://code.google.com/apis/socialgraph/
[*12] http://microformats.org/wiki/rel-tag

を直接反映させたメタデータを記述したときは情報探索に役立つが，そうでないときのメタデータは情報探索に役立たないという問題がある．

1.3.2 セマンティック Web

Tim Berners-Lee は，Web コンテンツの内容を機械が理解できるように記述したメタデータを生成して，情報探索や情報統合などの処理を自動的に行えるようにすることを目的としたセマンティック Web を提唱した [Berners-Lee 01]．セマンティック Web では，メタデータを記述する際，Resource Description Framework （RDF） [McBride 04] を使用することが提案されている．RDF は，主語 (Subject)，述語 (Predicate)，目的語 (Object) からなる，データ記述モデルである．Subject はリソース，Predicate はプロパティ，Object はプロパティの値を意味しており，RDF トリプルで表現する（図 1.3）．

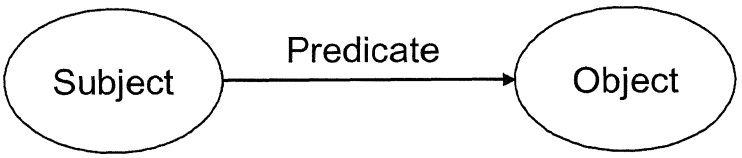

図 1.3: 基本の RDF トリプル

RDF トリプルは，Subject を URI 参照あるいは空白ノードで，Predicate を URI 参照で，Object を URI 参照あるいは文字列あるいは空白ノードで記述することができる．URI （Uniform Resource Identifier） [Tim Berners-Lee 05] 参照とは，一定の書式で書かれたリソース（電子ドキュメントや画像，人，会社，書籍など）の識別子である．例を図 1.4 に示す．

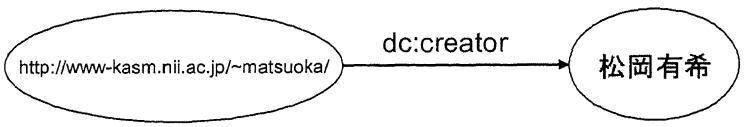

図 1.4: RDF トリプル例

図 1.4 は，`http://www-kasm.nii.ac.jp/~matsuoka/` を松岡有希が作成したということを示している．プロパティにあたる dc:creator は Dublin Core のリソースの作成者を表す要素である（表 1.1）．このデータモデルは下記に示すように XML 構文で記述することができる．

```
<? xml version="1.0" encoding="Shift_JIS" ?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/" >
    <rdf:Description rdf:about="http://www.kasm.nii.ac.jp/~matsuoka">
       <dc:creator>松岡有希</dc:creator>
    </rdf:Description>
</rdf:RDF>
```

プロパティを記述するときは，通常はRDFS（RDF Vocabulary Description Language 1.0: RDF Schema）[Brickley 04] で定義されたものを用いる．RDFSでは，基本クラスと基本プロパティが用意されている．メタデータを記述する人は，クラスを定義した上で，サブクラスやインスタンスを用いて知識を表現する．RDFSを利用することで簡単な知識を記述することはできるが，より詳細な知識を記述したいときはオントロジの記述言語であるOWL(Web Ontology Language)[McGuinness 04] を用いる．オントロジとは，知識を記述したり表現したりするために使用される語彙関係を定義したものである．OWLは複数のRDFトリプルから構成される．RDFSと比べて，より複雑なクラス関係を記述したり，定義されている制約条件を利用することでクラスの限定条件を記述することが可能になっている．RDF（S）やOWLを利用することで，ユーザはWebコンテンツの内容を機械が理解できるようなメタデータを記述することができる．オントロジを利用することによって，Webコンテンツの内の語の意味の定義を記述できるようになる．本研究において提案している意味的メタデータは，セマンティックWebが目標としているメタデータと同じである．

セマンティックWebの実現に向けて，RDF（S）やOWLで記述したメタデータを生成するためのツールがいくつか研究されている．

Annotea[Kahan 01] は，HTMLかXMLで書かれたWebコンテンツに，ユーザがRDFによるアノテーションを記述することができるフレームワークである．Annoteaを利用するにはAmaya[*13] というブラウザが必要である．また，Amayaアノテーションは MozillaやInternet Explorerのようなブラウザでも見ることができる．ユーザは，すでに用意されているRDFスキーマや自分で定義したRDFスキーマを用いて，Webコンテンツ内の好きな位置にメタデータを記述することができる．Annoteaでは，コンテンツ内のアノテー

[*13] http://www.w3.org/Amaya/

ションの位置を記述するために，Xpointer[Steve DeRose 01] を使用している．生成されたメタデータは，ユーザのローカルマシンか公式に用意されている RDF サーバに保存される．

OntoMat-Annotizer[Handschuh 03] は，ユーザが Web コンテンツに対して OWL を用いてメタデータを生成したり，メタデータを保守したりできるようなツールである．オントロジやインスタンスの探索や，テキスト内でアノテーションされたテキストを表示する HTML ブラウザといった機能が付いている．Java で作られており，拡張のためのプラグインを提供している．対象とするユーザは，Web コンテンツを OWL によるメタデータで高機能化したいと思っているユーザである．OntoMat では，ユーザが Web コンテンツ内の関連する部分をハイライトしたり，ドラッグアンドドロップで新しいインスタンスを作成できるようにしている．

Open Ontology Forge（OOF）[Collier 04] は，テキストや画像をオントロジに基づいてアノテーションするツールである．OOF では，ユーザが Web コンテンツ内の文字列をマウスで選択し，オントロジのクラスコンセプトにドラッグアンドドロップすることで，既存のオントロジと Web コンテンツ内の文字列とを手動でマッピングすることができる．この結果は RDF か XML 形式で保存することができる．

MnM[Vargas-Vera 02] はユーザがマークアップした情報を基に学習をして，半自動的にマークアップを行うツールである．メタデータは RDF などで保存される．

SemTag[Dill 03] は自動マークアップにだけ焦点をあてた自動アノテーションツールである．RDF で書かれたオントロジである TAP のコンセプトを，Web コンテンツ内の語に自動的にマークアップする．IBM のテキスト分析プラットフォーム Seeker を使用しており，曖昧性の問題に関しては，Taxonomy Based Disambiguation(TBD) アルゴリズムを提案している．

KIM[Popov 03] はセマンティックアノテーションを生成したり，検索を行うためのプラットフォームである．固有名認識（Named Entity Recognition）の技術を利用して，Web コンテンツ内の固有名詞とセマンティックレポジトリ内にあるオントロジのコンセプトをリンクする．コンテンツの中でも，統計分析から得られた特徴度の高い固有名詞を使用する．

このように，セマンティック Web におけるメタデータの生成方法は，手動で行うアプローチと自動で行うアプローチがある．手動でメタデータを生成するアプローチは，RDF や OWL，オントロジといった専門知識が必要である，また，ユーザがメタデータを作成するコストがかかるため，一般のユーザが手動によるシステムを利用するには敷居が高いと

いう問題がある．(半) 自動でメタデータを生成するアプローチの場合，ユーザのアノテーションのパターン分析や自然言語解析から自動的に生成されたメタデータは意味の定義が間違っている可能性があるという問題がある．また，これらのアプローチでは，Web コンテンツ内の主題語や特徴語がメタデータに記述されているかどうかは考慮されていない．

1.3.3 RSS と Atom

近年，WWW 上でもっとも普及しているメタデータは RSS であろう．RSS は通信社のニュースや個人の Weblog で用いられているメタデータのフォーマットである．RSS はいろいろなバージョン（表 1.3）があり，複数の規格が存在する [Pilgrim 02]．

RSS 0.9 と RSS 1.0 は RDF Site Summary の略であり，RSS0.91 は Rich Site Summary の，RSS2.0 は Really Simple Syndication の略である．表 1.3 から分かるように，UserLand が制定した RSS と RSS-DEV Working Group が制定した RSS とで大きく仕様が異なる．現在，日本では RSS1.0 が主に使われているが，世界的には RSS2.0 が広く使われている．また，RSS とは異なるフォーマットである Atom も普及している．Atom は RSS の考えを踏襲しながら，Weblog やニュースヘッドラインといったコンテンツの配信を目的としてできたフォーマットである．本稿では，RSS や Atom の書き方については詳しく述べないかわりに，どのような要素が規定されているかについて述べる．表 1.4 は，RSS1.0 と RSS2.0, Atom で規定されている要素の比較を示している．これによると，RSS2.0 は RSS1.0 に比べて，発行日時やクロールしてほしくない時刻などコンテンツ配信を目的とした要素が多いことが分かる．規定されていない要素は，Dublin Core モジュールを使って追加記述することができる．

これらの要素を 1.2 で紹介したメタデータの分類と比較すると，表 1.4 の description と表 1.1 の Description が Direct Content-based Metadata や Content-descriptive Metadata に相当し，それ以外の要素は Content Independent Metadata に相当する．description では，コンテンツの概要を書くように推奨されているが，実際はコンテンツに書かれている文章がそのまま記述されていることが多い．これは RSS や Atom をコンテンツの配信目的に使用していることと，これらのメタデータは機械によって自動的に生成されているからである．RSS1.0 では，RDF トリプルを利用してメタデータを記述できるので，意味的メタデータを生成することは可能である．しかし，現状では他の RSS2.0 や Atom と同じく，タイトルや著者，本文がそのまま書かれたものが大半である．

表 1.3: RSS のバージョン

バージョン	制定者	長所	現状	推薦ポイント
0.90	Netscape		1.0 の登場により使われていない	使用しないでください
0.91	UserLand	とても簡単	公式には 2.0 の登場により廃れたことになっているが，まだ使われている	必要最低限の要素を書く為に使ってください．もし，もっと柔軟に書きたいのであれば 2.0 を使ってください．
0.92, 0.93, 0.94	UserLand	0.91 よりも複雑なメタデータを書くことができる	2.0 の登場により使われていない	このバージョンの代わりに 2.0 を使ってください
1.0	RSS-DEV Working Group	RDF ベースなので，モジュールを使って拡張ができるし，一つのベンダーによってコントロールされているわけではない．	コア部分が決定されている．モジュール部分は開発段階である．	RDF ベースのアプリケーションを使ってください．あるいは高度な RDF のモジュールが必要なら使ってください．
2.0	UserLand	モジュールによる拡張性がある，バージョン 0.9X から簡単に移行できる	コア部分が決定されている．モジュール部分は開発段階である．	一般的な目的や高度なメタデータを書くために使ってください．

表 1.4: RSS1.0,2.0 および Atom の規定要素

説明	RSS1.0	RSS2.0	Atom1.0
タイトル	title	title	title
配信サイト or 記事の URI	link	link	link
概要	description	description	subtitle,summary and/or content
記述言語	-	language	-
著作権表示	-	copylight	rights
技術的担当者	-	webMaster	-
配信担当者	-	managingEditor	-
発行日時	-	pubDate	published
更新の最終日時	-	lastBuildDate	-
カテゴリー	-	category	category
プログラム名	-	generator	generator
フォーマット名	-	docs	-
更新の通知	-	cloud	-
有効期限	-	ttl	-
ロゴ画像	image	image	logo, icon
クロールしてほしくない時刻	-	skipHours	-
クロールしてほしくない曜日	-	skipDays	-
配信記事	item	item	entry
著者名	-	author	author
配信記事に貢献した人や実体	-	-	contributor
コメントを受け付けている URL	-	comments	-
添付メディアファイル	-	enclosure	-

1.3.4 ソーシャルブックマーク

ソーシャルブックマークは，ユーザがタグと呼ばれるキーワードと共に Web コンテンツをブックマークし，複数のユーザ間でブックマーク情報を共有するサービスである．タグは，ユーザが Web コンテンツを整理したり，思い出しやすくするために Web コンテンツに与えるキーワードによる説明である．ユーザは，自由な言葉を使ってタグを付与したり，一つの Web コンテンツに対して複数のタグを付与することができる．代表的なサービスは，del.icio.us[*14] やはてなブックマーク[*15] などである．ユーザはタグを通じて，タグに関連する情報を獲得することができる．また，簡単なので誰でも参加できることが利点である．[Golder 06] によって，タグは下記のように分類されている．

(1) Web コンテンツの主題に関すること

(2) Web コンテンツに書かれている内容の種類
　例：article, blog, book

(3) Web コンテンツを作成した人の名前

(4) 単独では意味がなく，分類のためのタグ
　例：丸めた数字，記号

(5) タグを付与したユーザの意見を反映した形容詞
　例：scary, funny, stupid

(6) Web コンテンツとタグを付与したユーザの関係
　例：mystuff, mycomments

(7) Web コンテンツに対するユーザのタスク
　例：toread, jobsearch

(1)〜(3) のようなタグは Web コンテンツの内容と直接関係するタグのため，ユーザがこれらのタグにブックマークされている Web コンテンツを見たとき，タグの内容に即した Web コンテンツを取得できる．ユーザは (1)〜(3) のようなタグが付与されているコンテンツの中からほしい情報を探すことができるが，タグの中にはユーザが意図していた意味

[*14] http://del.icio.us/
[*15] http://b.hatena.ne.jp/

とは異なって利用されている場合があり，情報探索の手間がかかる．また，ユーザはタグを通して，他のユーザが付与しているコンテンツの情報を得ることによって，ユーザが探している情報と関連のある情報を探すことができる．ソーシャルブックマークでは，(4)〜(7) のような個人的な意見や解釈が反映されたタグがブックマークされているWebコンテンツを見ても，タグを付与したユーザ以外は期待通りの情報を獲得しにくいという問題がある [Mathes 04]．例えば，funnyやtoreadといったタグはWebコンテンツに対する評価や重要度がユーザによって異なるので，これらのタグにブックマークされているWebコンテンツを見ても役に立たないユーザがいる．このように，ソーシャルブックマークにおいてユーザが自由な言葉で付与したタグの中には，タグの内容に即した情報を取得したい場合に適していないものもある．

ソーシャルブックマックのタグから意味的メタデータを生成する研究がいくつか行われている．大きく分けて2種類のアプローチがある．一つは，ソーシャルブックマークのタグからオントロジを生成する，というアプローチである．この場合，ソーシャルブックマークのタグがオントロジのコンセプトになる．[Mika 05] や [Xian Wu 06] らは，ソーシャルブックマークのタグとタグを付与したユーザ，タグが付与されているWebコンテンツの3つの関係を使って，タグ間の関連性の発見を試みている．Mikaは，タグ間の上位・下位概念を見つけることによって，ライトウェイトオントロジを作った．Wuは，概念が似ている語集合を見つけ出した．

もう一つのアプローチとして，既存のオントロジのコンセプトとソーシャルブックマークのタグとをマッピングすることによってメタデータを生成する，というアプローチがある．Specia[Specia 07] やDamme[Damme 07] らは，オントロジを利用してタグに意味を付加するというアプローチをとっている．Speciaは，既存のオントロジのコンセプトやプロパティやインスタンスにタグをマッピングしたり，マッピングされたタグ間の関係を決定したりする．Dammeは，ソーシャルブックマークのタグだけでなく，WordNet[*16] やWikipedia[*17] といった辞書やRDFやOWLで書かれたオントロジを利用して，タグとそれらをマッピングすることでメタデータの生成を試みている．

[*16] http://wordnet.princeton.edu/
[*17] http://wikipedia.org/

1.4 まとめ

本研究では，ユーザがほしい情報を獲得しやすくするために，Web コンテンツの主題を表す語や特徴語，コンテンツの内容に関連する語，コンテンツ内の語の意味の定義などが記述されている意味的メタデータの生成を目標としている．既存のメタデータを検証した結果，下記のような問題があることが分かった．

- Web コンテンツの著者がメタデータを生成する場合，Web コンテンツの内容と関係のないことが意図的に書かれる可能性がある
- 専門的な知識が必要で，ユーザコストがかかるシステムは，一般ユーザが利用するには敷居が高い
- 自動的に生成されるメタデータは，Web コンテンツの内容を代表するメタデータが生成されるとは限らない
- ソーシャルブックマークサービスにおいて，Web コンテンツのタグ付けは容易なので一般ユーザでも生成することはできるが，ユーザの主観が反映されたタグが付与されることがある

次章で，これらの問題の解決に向けたアプローチについて述べる．

1.5 本論文の構成

本論文は 6 章から構成される．第 1 章は，本研究の背景や目的，既存のメタデータ生成における問題点について述べた．第 2 章では，本研究で想定している意味的メタデータを生成するためのアプローチや研究手法について述べる．第 3 章では，ユーザは下線引きによってページ内のどのような箇所に着目するのかについて，イロノミーというシステムで得られたデータを基に行った分析について述べる．第 4 章では，ユーザは他人が付与したマーキングデータを用いたページ探索を好むかについて，合口というシステムで得られたデータを基に行った分析について述べる．第 5 章では，ユーザがコンテンツに付与したメモは内容と関連があるのかについて，memoQ というシステムで得られたデータ基に行った分析について述べる．これらの成果について，第 6 章でまとめ，本論文を結ぶ．

第 2 章
本研究におけるアプローチ

1.3 で述べたように，既存のメタデータは，本研究で目標としている，ユーザが情報を探しやすくなるような意味的なメタデータを生成するには問題があった．そこで本章では，メタデータを誰が，どのように生成すれば問題を解決できるのかという観点から，本研究のアプローチを述べる．その後，メタデータ生成に向けての全体像および，具体的な研究手法について述べる．

2.1 誰がメタデータを生成すればよいのか

HTML の META タグでは，メタデータを Web コンテンツの著者が生成すると，Web コンテンツの内容と関係のないことが書かれるおそれがあるという問題があった．一方で，ソーシャルブックマークタグのように Web コンテンツの読者がメタデータを生成すると，主観が入るおそれがあるという問題があった．では，Web コンテンツの著者と読者のどちらがメタデータを生成すべきだろうか．Web コンテンツの著者がメタデータを生成する場合，著者一人の意見が反映されるだけである．一方で，Web コンテンツの読者がメタデータを生成する場合，一つの Web コンテンツにつき，複数人の知識を利用してメタデータを生成することが可能である．複数人でメタデータを生成することによる利点は，集合知を利用できることである．集合知に関する著書，「The Wisdom of Crowds」[Surowiecki 05] では，“多様性”，“独立性”，“分散性”，“集約性” が機能するときに，集団の知恵を活かすことができるとしている．これらの要件の説明は下記のとおりである．

多様性 それが既知の事実のかなり突拍子もない解釈だとしても，各人が独自の私的情報を多少なりとも持っている

独立性 他者の考えに左右されない

分散性 身近な情報に特化し，それを利用できる

集約 個々人の判断を集計して集団として一つの判断に集約するメカニズムの存在

集合知では，多様で自立した個人から構成される，ある程度の集団に予測や推測をしてもらい，集団の回答を集約すると，個人が回答を出す過程で犯した間違いが相殺される，と言われている．本研究では，Web コンテンツの内容を直接反映した語が記述された意味的なメタデータを生成することを目標としている．一人のユーザのみがメタデータを作成すると，ユーザの主観が反映されたメタデータが出来上がるかもしれない．しかし，集合知が機能する仕組みをユーザに提供することができれば，複数ユーザの知識を集約することによって，Web コンテンツの内容を代表するメタデータを生成できる可能性がある．したがって，本研究では，Web コンテンツの複数の読者の知識を意味的メタデータの作成に利用することを提案する．

2.2 どうやってメタデータを生成すればよいのか

セマンティック Web におけるメタデータ生成に関する先行研究によると，Web コンテンツ内の語の意味の定義に関連する語をユーザに記述してもらおうとすると，ユーザに多大な負担をかけてしまう，という問題があった．その点，ソーシャルブックマークでは，一般ユーザが自身の視点で自由にタグを付与することができるため，ユーザにかかる負担は少ない．しかし，この場合，Web コテンツの内容とは関係のないタグを書かれてしまう場合があるという問題がある．したがって，大勢のユーザにメタデータを生成してもらうためには，ユーザにとって負担が少なく，Web コンテンツの内容が反映されたメタデータが付与されるようにする仕組みが必要となる．

そこで本研究では，人が日常生活において本や文献などの文章を読む際，下線を引いたり，メモ書きをするアノテーション行為に着目した．紙ベースの文章に対するアノテーションにどのような形式および効果があるのかについての調査によると [Marshall 97]，教科書に行ったアノテーションの形式および機能については，表 2.1 のような結果が得られている．これにより，ハイライトや下線引きは，ユーザが後で読み返すため，あるいは，記憶するために使われているため，ユーザは Web コンテンツの内容を代表する箇所に着目している可能性が高い．また，メモはユーザの解釈を記述するために使われていることから，メモには Web コンテンツの内容と関連する語が含まれる可能性が高い．したがって，ユーザがアノテーションを付与したデータからは，本研究で目標としている意味的なメタデータに有用な語を獲得できるものと思われる．また，アノテーション行為は，ユーザが自身

のために行う行為であるため，ユーザに負担をかけることはない．

表 2.1: アノテーションの形式および機能

形式	機能
ハイライトや下線引き，アスタリスクなどの記号，×印で消された文字	後で気付かせるための目印
短いハイライト，単語やフレーズの囲み，テキスト内のマーキング，アスタリスクのような余白のマーキング	記憶するため
余白内の，あるいは図の近くの，あるいは質問にたいする注釈	問題解決のためのアノテーション
余白の注釈，テキスト間の長いメモ，行間のワードやフレーズ	ユーザの解釈
(拡張された) ハイライトや下線	難しい文章や長い文書のとき，読者の注意の可視化された足跡
メモ，絵，素材と関係のないマーキング	読書中の偶然の反応

さらに，アノテーションを利用することによって，情報価値の高い Web コンテンツにだけ意味的メタデータを生成できる可能性がある．現在の WWW 上では，Splog と呼ばれるスパムコンテンツを生成する Weblog により，情報価値のないコンテンツが大量に生成されている．Splog は link spam blogs と fake blogs の二種類がある [Kolari 06]．link spam blogs は，広告サイトへのリンクが大量に記述された blog のことで，fake blogs は WWW 上にある Web コンテンツから不当にテキストを抽出することによって作成された blog のことである．情報探索をする際，最も利用されている検索エンジンでは，検索結果に Splog のようなスパムコンテンツが含まれることが問題となっている．これはコンテンツに対して自動的にインデックス（メタデータ）を生成していることが原因の一つであると考えられる．機械がコンテンツに情報価値があるかどうかを判断するためには，コンテンツに記述されている文字列のパターンを学習することが挙げられる．しかし，日々，様々な文字列パターンのコンテンツが大量に生成されているため，全てのスパムコンテンツを排除するのは難しい．ユーザのアノテーション行為を利用すれば，スパムコンテンツにメタデータを生成されるということは少なくなるだろう．逆に，情報価値の高いコンテンツに対してメタデータが生成されるようになることが期待できる．

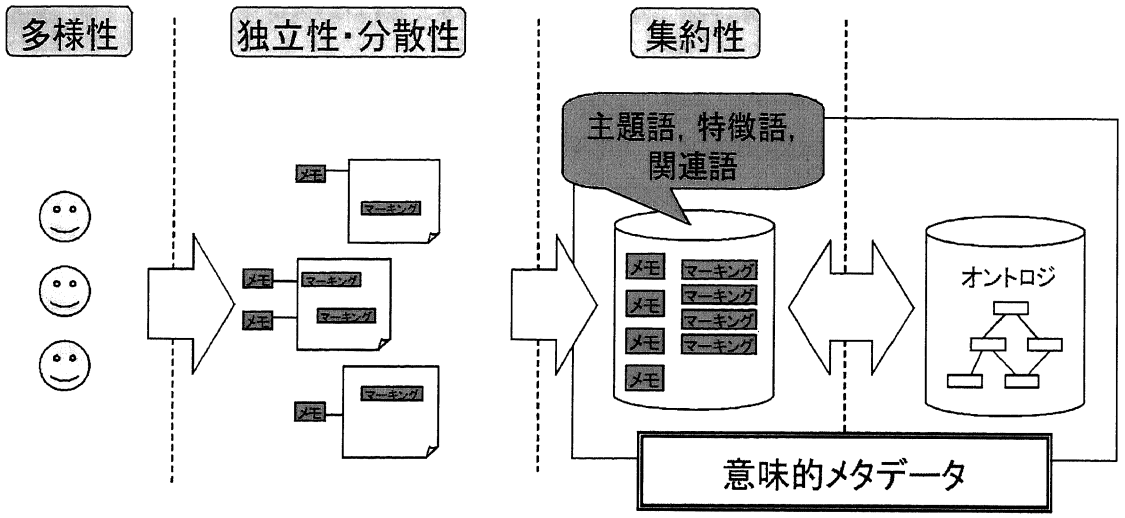

図 2.1: 意味的メタデータ生成に向けた全体像

2.3 意味的メタデータ生成に向けての全体像

本節では，意味的メタデータ生成に向けた全体像について述べる．図 2.1 に本研究で想定している，メタデータ生成に向けての全体像について示した．まず初めに，多様な意見を持った複数のユーザに，Web コンテンツに対してマーキングやメモ書きといったアノテーションを付与してもらう．ユーザがアノテーションを付与するときは，他人の意見に左右されないような独立性や，身近な情報に特化できるような分散性といった環境が用意されていることが望ましい．次に，複数ユーザが付与したアノテーション情報を集約することによって，Web コンテンツの主題語や特徴語，コンテンツの内容と関連する語を獲得する．さらに，既存のオントロジを利用することでアノテーションから獲得した語の意味の定義を行うことによって，意味的メタデータが生成できると想定している．

2.4 研究手法

本節では，研究ポイントと実験環境について述べる．

2.4.1 研究ポイント

WWW 上にある大量の Web コンテンツのために意味的なメタデータを生成するには，誰でも参加できるアーキテクチャが必要である．そこで，本研究では，人が日常的に行うアノテーションに注目した．アノテーションに関する先行研究では [Marshall 97]，ユーザがどういう目的でアノテーションを行ったのかについて調査しているが，アノテーションされた箇所はどういった語なのかといった調査はされていない．本研究では，複数ユーザが付与したアノテーションから意味的なメタデータを作成するために有用な語を獲得できるのかについて調べることを目的としており，図 2.1 で示した，アノテーションから獲得した語とオントロジのマッピングに関する研究提案は行っていない．

分析では，Web コンテンツの内外に含まれる語から，意味的メタデータに有用な語を獲得できるのかについて調べる．Web コンテンツ内の語から意味的メタデータを生成する方法としては，下線引きやマーキングに着目した．下線引きやマーキングは，ユーザがコンテンツ内の語を選択することと同意なので，主観が反映された語は含まれないからである．ユーザが下線やマーキングを付与したデータの分析ポイントは，下記のとおりである．

分析 1 下線が付与された文字列からコンテンツの主題を表す語や特徴語を獲得できるのか?

分析 2 マーキングが付与された語や文字列はユーザ間で共有した場合に情報探索に役立つのか？

Web コンテンツ外の語からの意味的メタデータの生成方法としては，メモ書きに着目した．コンテンツに含まれない語でもコンテンツの内容と関連する語を獲得できる可能性があるからである．ユーザがコンテンツに付与したメモの分析ポイントは下記のとおりである．

分析 3 コンテンツの内容と関連のある語が獲得できるのか？

これら 3 種類の分析を通じて，複数ユーザによって付与され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のかについて調査した．

2.4.2 実験環境

本研究では，2.4.1 で述べた 3 種類の分析を行うために，2005 年〜2007 年の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されたアノテーションシス

テムの運用で得られたデータを使用した．本節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について紹介する．

人工知能学会全国大会では，2003 年から 2007 年にかけてイベント空間情報支援プロジェクトを行ってきた [西村 04, 武田 06]．このプロジェクトはイベントが開催される実空間において参加者・主催者双方の満足度の向上を目的としたもので，オープンな共通プラットフォームを構築しながら，産学官共同で有用な情報支援システムを社会に提供することを目指している．大会支援システムはそのひとつの具体例である．大会支援システムは，会場支援システムと Web 支援システムからなる．

本研究で利用した Web 支援システムは，コンテンツ技術，コミュニティ技術を利用しており，オンライン上での利用者間のコミュニケーションや情報共有の促進，さらにはソーシャルウェア的な機能による学会情報への容易なアクセスを提供している．図 2.2 は JSAI2005[*1] の，図 2.3 は JSAI2006[*2] の，図 2.4 は JSAI2007[*3] の大会支援システムのトップページである．

図 2.2: JSAI2005 大会支援 Web システムのトップページ

[*1] http://jsai-support-wg.org/polysuke2005/
[*2] http://2006.jsai-support-wg.org/
[*3] http://jsai2007.polypho.net/

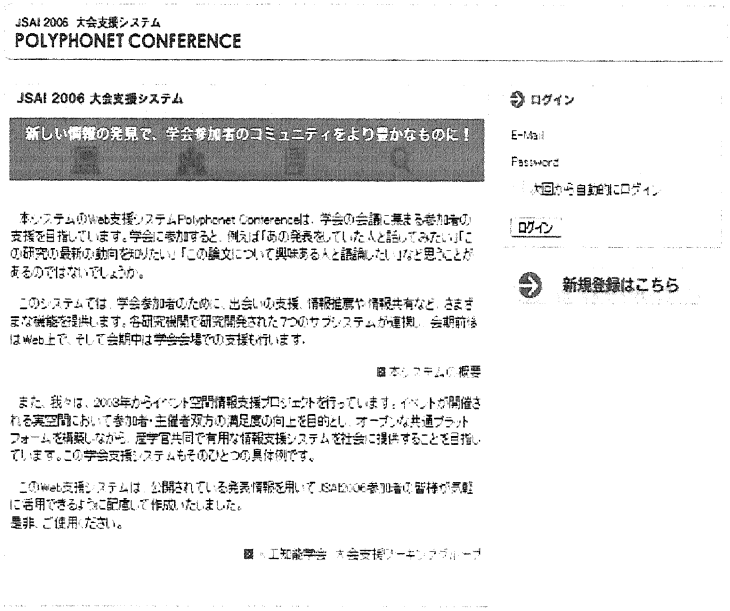

図 2.3: JSAI2006 大会支援 Web システムのトップページ

図 2.4: JSAI2007 大会支援 Web システムのトップページ

Web 支援システムでは，ユーザの名前や所属，発表情報が書かれた My ページを用意している．学会参加者はユーザ ID およびパスワードを使って Web 支援システムにログインすると，My ページにアクセスすることができる（図 2.5）．また，知り合いの研究者を友人登録することができるようにもなっており，既存の SNS サービスと同等の機能が提供されている．その他の機能としては，学会のタイムスケジュールや発表論文ごとに論文の概要や著者名が書かれた発表ページが用意されている（図 2.6）．本研究では，この発表ページを対象としたアノテーションシステムを運用することによって得られたデータを使って分析を行った．分析 1 では 2005 年に運用された “イロノミー” を，分析 2 では 2006 年に運用された “合口” を，分析 3 では 2007 年に運用された “memoQ” で得られたデータを使って分析をした．

図 2.5: JSAI2007 大会支援 Web システムの My ページ画面

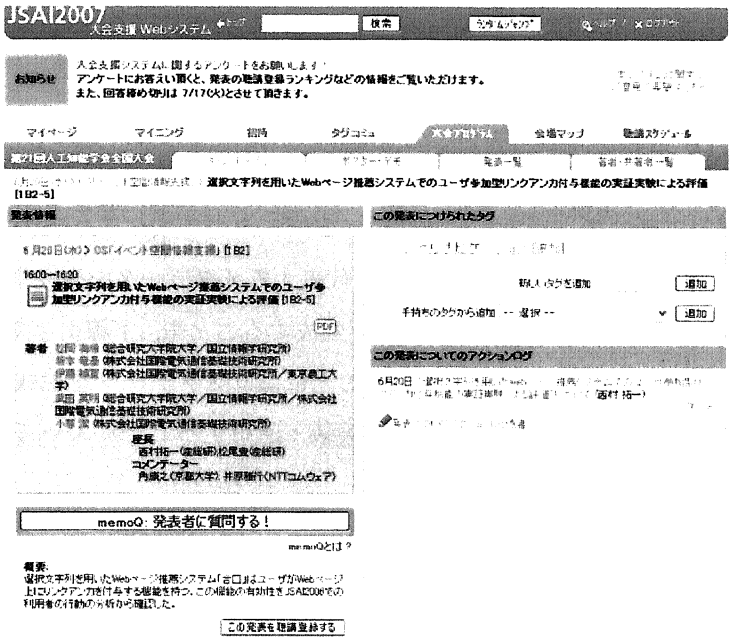

図 2.6: JSAI2007 大会支援 Web システムの発表ページ画面

2.5 関連研究

複数ユーザが Web コンテンツに対してアノテーションを付与できるシステムは，これまでにたくさん開発されている．これは WWW 上で任意の Web コンテンツに対してユーザが意見を書き込んだり，ユーザ同士で意見交換をしたりと，ページの作者とその読者，あるいは読者間同士といった双方向による知識共有の需要が高まったためである．

iMarkup[*4] は，Web ブラウザを通じて Web コンテンツにマウスで絵を描いたり付箋を貼り付けることができる．一方で，iMarkup のように自由に描画できるアノテーション機能はなく，コメントのみを付与することができるシステムが多々ある．ComMentor[Röscheisen 94] や CoNote[Davis 95]，CritLink[Yee 02]，YAWAS[Denoue 00] は，Web ブラウザを通して，Web コンテンツの一部分を特定し，付箋のようにコメントを付与できる．なかには，コメントに対してコメントを追加することで，オリジナルの Web コンテンツ上でスレッド形式でディスカッションすることができる．このように，ユーザ間のコミュニケーションを目的としたアノテーション共有システムは多々ある．しかし，アノテーションによってユー

[*4] http://www.imarkup.com/

ザが注目した箇所にどういった特徴があるのか，といったアノテーションのデータを使った分析研究は少ない．

2.6 まとめ

本章では，情報探索に役立つような意味的メタデータを生成するためのアプローチについて述べた．本研究では，コンテンツの内容を代表するメタデータを生成するために，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や，下線引きやメモ書きといったアノテーションから得られた語を利用することによりメタデータを生成することを提案した．次章以降，アノテーションシステムを実装・運用することによって得られたデータを使って各種分析を行った．

第 3 章

複数ユーザが付与した下線文に関する分析

本研究では，Webコンテンツの内容が直接反映された語が記述された意味的メタデータを作成するために，ユーザが文章を読むときに下線を引く行為に着目した．本章では，複数ユーザがWebコンテンツに付与した下線にどのような特徴があるのかについて調査した．調査に利用したのは，第19回人工知能全国大会（JSAI2005）で運用された，三色ボールペン読書法に基づいて下線を引くことができるシステム“イロノミー”である．本章では，イロノミーの運用で得られたデータを使って行った分析について述べる．

3.1 はじめに

本研究では，メタデータにWebコンテンツの内容が直接反映された語が書かれていれば，情報探索に役立つと考えている．たとえば，「macにインストールすべきソフトウェア」と題したブログ記事にたいして，「あとで読む」という語が記述されたメタデータがあったとしても，macに関心のない人やmacを熟知している人にとっては「あとで読む」という情報は意味をなさない．もし，「mac」という語がメタデータとして記述されていれば，このブログ記事はmacに関する情報が書かれているということが分かるため，誰にとっても有益な情報となる．

そこで本研究では，Webコンテンツの内容が直接反映された語を獲得するために，ユーザのアノテーション行為の一つである下線引きに着目した．下線を引くという行為は，文章内の語を選択することであるから，下線が付与された語はコンテンツの内容が直接反映された語といえる．したがって，下線が付与された箇所には，コンテンツの内容と関係のない語は含まれないと考えられる．また，[Marshall 97]では，大学の教科書におけるアノテーション行為を調べた結果，人は他で引用するために重要な文章を記録したり，後で見直したりするために下線を引き，線の色は情報の種類をコード化するために用いると報告

している．さらに，[Gyunn 78] では，学習目的で文章を読む場合，下線引きは読み手が重要だと考える情報を探し出す探索・選択過程であるといえる，と報告している．このように，ユーザはコンテンツ内の重要箇所に下線を引く，ということが報告されていることから，下線が付与された語を意味的メタデータとして活用できる可能性がある．本節では，Web コンテンツに下線を引くことができるシステムの運用から得られたデータを基に，複数ユーザが付与した下線のデータを集約すると，どのような語が獲得できるのかについて定量的指標を用いて調査した．

3.1.1 関連研究

[Blanchard 87] は，探索・選択過程は個人によって異なるものであり，また慣用的なものであるため，原文の文章を理解して下線を引く人もいれば，あらかじめ持っている知識を基に下線を簡単に引く人もいる，と述べている．また，[Morris 77] は，重要箇所の探索・選択に成功するかどうかは，下線を引いているときの認知的処理の深さおよび量によるといえる，と述べている．したがって，下線が付与された箇所が必ずコンテンツ内の重要箇所である，という保障はない．本研究では，複数ユーザによって付与された下線データを集約することで，意味的メタデータに有用な語の獲得を目指す．

3.2 イロノミー

本研究では，ユーザが下線を付与した語のデータを取得するために，JSAI2005 で運用された “イロノミー” を利用した．イロノミーは，[坂本 06] が開発した，大会支援システム内にある発表ページ (学会で発表される論文の情報が書かれたページ) 内の論文概要の文章に対し，三色ボールペン読書法 [齋藤 03] に従ってユーザが色付きの下線を付与できるシステムである．三色ボールペン読書法は，客観的にとても重要だと思う箇所は赤色で，客観的にまあ重要だと思う箇所は青色で，主観的に重要だと思う箇所は緑色で下線を引きながら読書をする方法である．イロノミーでは，この三色ボールペン読書法と同じように，客観的または主観的に赤・青・緑の三色を使って下線を引くことができるようになっている．

システムの使用者が，論文概要文に下線を引くときの操作は下記のとおりである．

1. 大会支援システムにログインして，発表ページを表示する．
2. 下線を引きたい箇所をマウスカーソルでなぞり，指定されたボタンをクリックする (図 3.1)

3. 下線の色を選択して，“色線を引く” というボタンをクリックする (図 3.2)

4. 色線が引かれた画面が表示される（図 3.3）

イロノミーでは，同じ箇所に複数の色の下線を付与することができない．これは HTML の仕様上不可能だからである．

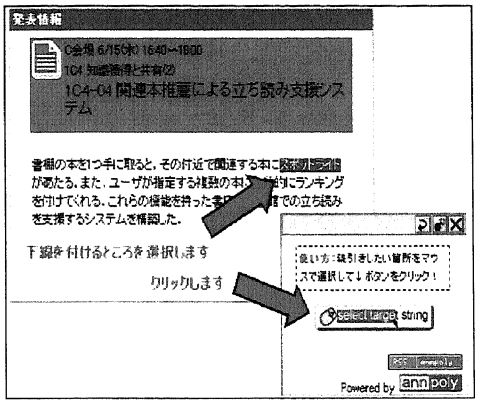

図 3.1: 下線を引く箇所を選択する画面

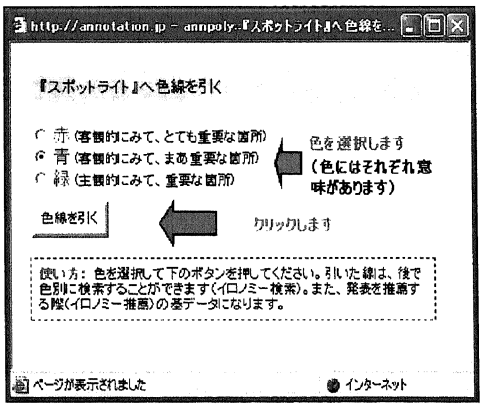

図 3.2: 下線の色を選択する画面

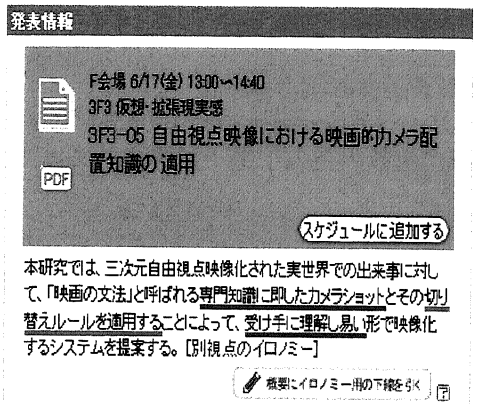

図 3.3: 下線が付与された発表ページの画面

3.2.1 運用結果

イロノミーが対象とした Web ページは，学会で発表される論文の概要文が書かれた発表ページで，全部で 294 ページあった．運用の結果，イロノミーを使用したのは開発者を除いて 27 人だった．ユーザが付与した下線の数の平均本数は 6.2 本で，分散は 87.2，標準偏差は 9.3 だった．分散の値が大きいのは，44 本の下線を付与したユーザがいれば，たった 1 本だけ下線を付与したユーザもいたからである．下線が付与された論文概要は 67 個あり，下線の総数は 168 本で，赤線の数が 47 本，青線の数が 64 本，緑線の数が 57 本だった．論文概要はそれぞれ 50 文字から 250 文字の間で書かれており，平均文字数は 140 文字だった．下線が付与された文字列はそれぞれ 2 文字から 33 文字であり，平均文字数は 10 文字だった．今回の実験は，被験者の属性の制御や統制が不十分な環境で行っている．また，全被験者がどの程度三色ボールペン読書法を理解してマーキングをしたかも不明であり普遍性があるデータとはいえないが，一つの傾向として報告する．

システムの運用によって得られた下線に関するデータは，以下のとおりである．

- 下線が引かれた日付
- 下線を引いたユーザ ID
- 下線を引かれた論文概要 ID
- 下線の色
- 下線が引かれた文章

分析で使用できるのは，上記のデータと論文概要文のデータである．

3.3 分析

本節では，ユーザがWebページ内で下線を付与した語を集約すると，どのような語を獲得できるのかについて，イロノミーの運用で得られたデータを使って分析を行う．分析対象として論文概要および下線文における名詞と未知語に注目した．形態素解析には茶筌[松本 03]を利用した．名詞を採用する理由は，下線が引かれた文章内で一番多く使われていた品詞だからである(表3.1)．また，未知語と判断された語は名詞のため，未知語も分析対象として扱う．分析では，294個の論文概要の文書(3個の論文概要は英文のため排除した)と下線が引かれた文章から抽出した名詞と未知語を利用する．

表3.1: 下線文に含まれる品詞

	名詞	助詞	動詞	助動詞	未知語	副詞
赤線	140	4	20	0	2	4
青線	172	8	24	1	6	4
緑線	141	7	20	1	3	5
全下線	453	19	64	2	11	13

3.3.1 コンテンツ内の語と下線が付与された語の比較

ここでは，コンテンツ内の語と下線が付与された語の比較を行うことで，下線文にコンテンツ内のどのような語が含まれているのかについて調べた．分析には，文書内の語を定量的に特徴付ける手法として広く用いられているtfidf[Salton 91]を用いた．文書内の語は，tfidfで求めた値を使うことにって，下記のように特徴付けることができる．

- tfidf値が高い語は，対象文書内で出現頻度が高く，他の文書には現れにくいので，対象文書の特徴語と言える．
- tfidf値が低い語は，対象文書内での出現頻度が低く，他の文書に頻繁に出てくる語であるため，一般語である可能性が高い．

分析手法を図3.4にまとめた．

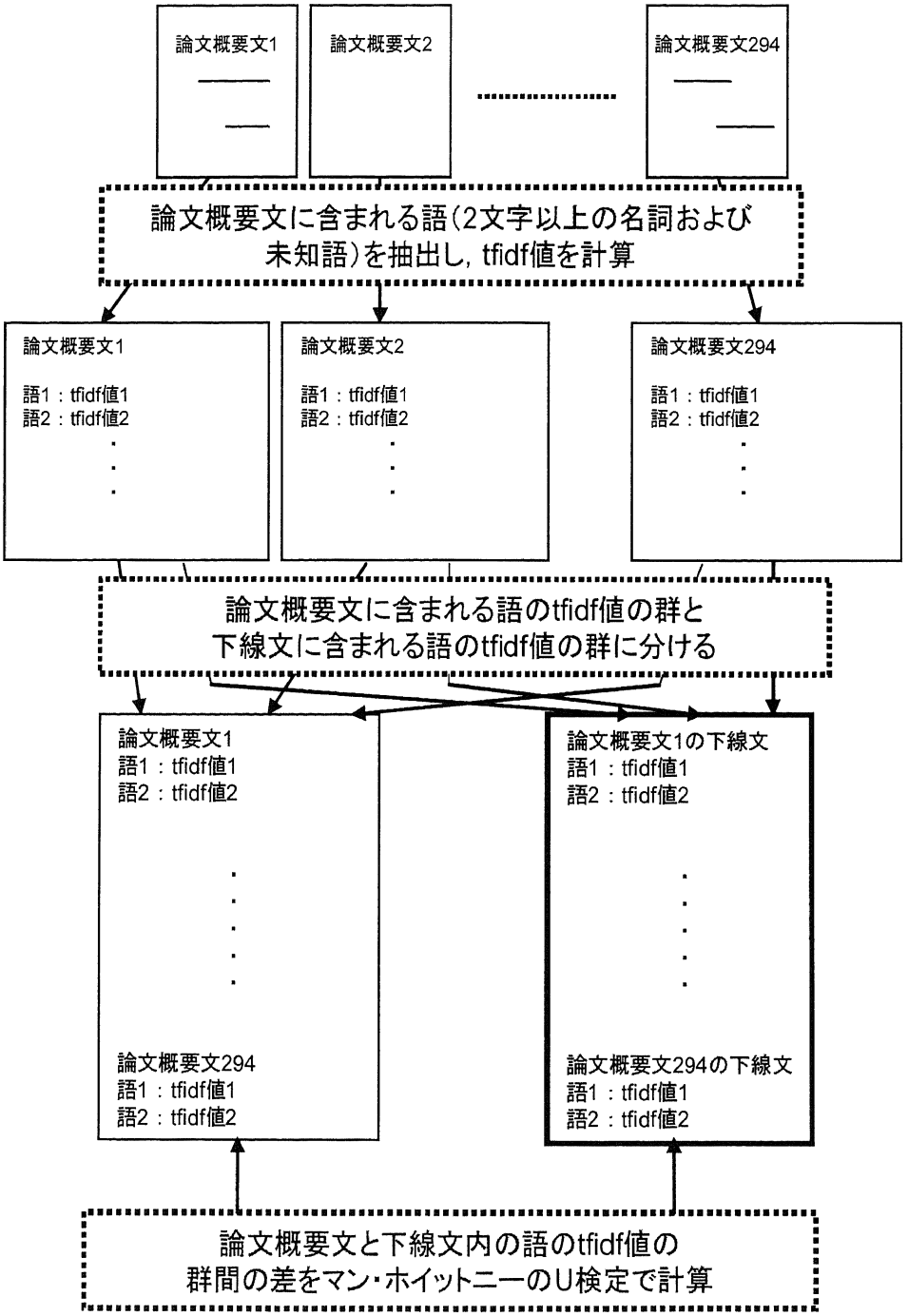
論文概要文1
論文概要文2
論文概要文294
論文概要文に含まれる語(2文字以上の名詞および未知語)を抽出し, tfidf値を計算
論文概要文1
語1 : tfidf値1
語2 : tfidf値2
論文概要文2
語1 : tfidf値1
語2 : tfidf値2
論文概要文294
語1 : tfidf値1
語2 : tfidf値2
論文概要文に含まれる語のtfidf値の群と下線文に含まれる語のtfidf値の群に分ける
論文概要文1
語1 : tfidf値1
語2 : tfidf値2
論文概要文294
語1 : tfidf値1
語2 : tfidf値2
論文概要文1の下線文
語1 : tfidf値1
語2 : tfidf値2
論文概要文294の下線文
語1 : tfidf値1
語2 : tfidf値2
論文概要文と下線文内の語のtfidf値の群間の差をマン・ホイットニーのU検定で計算

図 3.4: 分析手法

論文概要文に含まれる語は，各論文概要文ごとに茶筌を使って形態素解析をし，2文字以上の名詞および未知語を採用した．tfidf値は，この語を使って式3.1により求めた．

$$tf\,idf(w,a) = tf(w,a) \cdot idf(w) \tag{3.1}$$

$tf(w,a)$ は論文概要文 a における語 w の出現回数を，$idf(w)$ は 語 w が全論文概要文のうちどのくらいの頻度で出現するかの尺度であり，$\log(N/\,df(w))$ で求める．$df(w)$ は語 w が含まれる論文概要文の数を表す．N は論文概要文の総個数を表し，今回は294個である．下線文に含まれる語は，下線文に元の論文概要文内の語がある場合，その語を下線文に含まれる語として採用した．また，下線文に含まれる語のtfidf値は，元の論文概要文に含まれる語のtfidf値をそのまま利用した．表3.2は全ての論文概要文と全ユーザによって付与された下線文に含まれる語のtfidf値群の中央値および四分位偏差値，語数を示している．

表 3.2: 全論文概要文と全下線文に含まれる語に関するtfidf値

	全論文概要文	下線文
中央値	4.1	4.6
四分位偏差値	3.1	3.0
語数	6481	456

まず，全論文概要文に含まれる語群のtfidf値と全ユーザによって付与された下線文に含まれる語群のtfidf値の分散に差がないかどうかについて調べた．アンサリ・ブラドレイ検定を行った結果，$p > 0.05$ となり，両群の分散に有意差がないということが分かった．次に，全論文概要文に含まれる語群のtfidf値と全ユーザによって付与された下線文に含まれる語群のtfidf値とで中央値の差がないかを調べるために，マンホイットニーの U 検定を行った．検定の結果，U=1144928($p < 0.05$) となり，全論文概要文に含まれる語群のtfidf値と下線文に含まれる語群のtfidf値の中央値の差が有意であるという結果が得られた．図3.5は全論文概要文に含まれる語のtfidf値のヒストグラムで，図3.6は全下線文に含まれる語のtfidf値のヒストグラムである．図3.5, 3.6において，tfidf値のデータ区間が6以降の語の相対度数を見比べると，論文概要文に含まれる語の相対度数よりも下線文に含まれる語の相対度数のほうが高いことが分かる．従って，複数ユーザが付与した下線文を集約すると，tfidf値の高い語，すなわちWebページ内の特徴語を多く含む傾向があることが分かった．

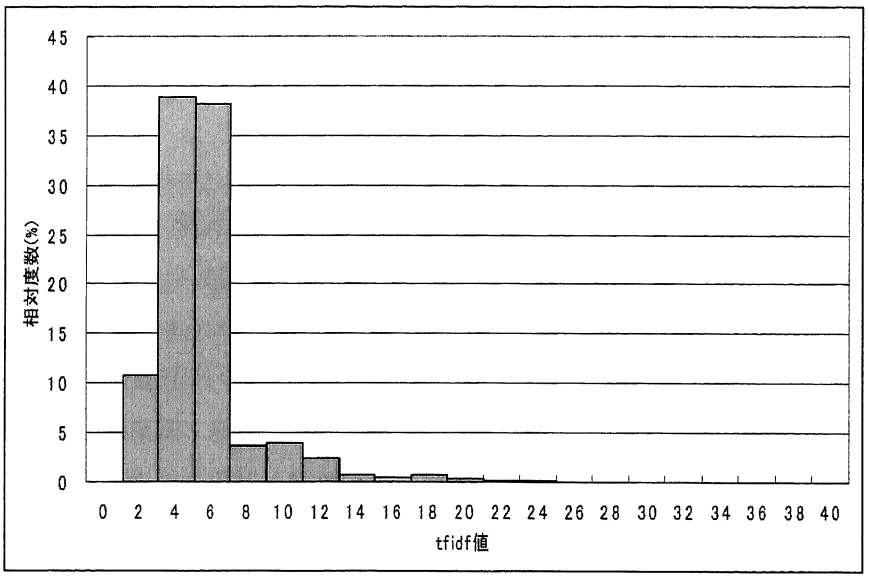

図 3.5: 全論文概要文に含まれる語の tfidf 値のヒストグラム

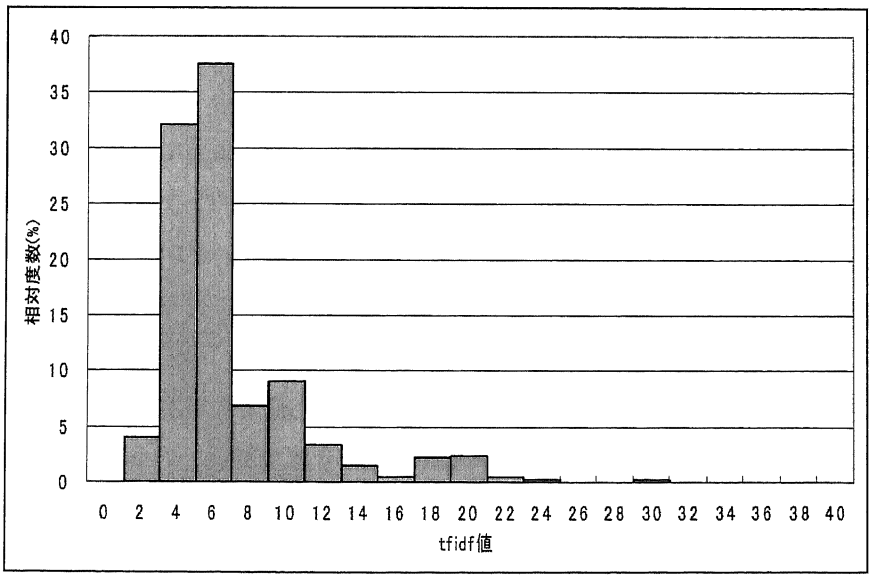

図 3.6: 全下線文に含まれる語の tfidf 値のヒストグラム

3.3.2 色線の比較

ここでは，全ユーザによって付与された色線のうち，各色線に含まれる語の特徴度に違いがあるのかを調べるために，tfidfを利用して分析を行った．図3.7は，赤・青・緑色の下線文に含まれる語のtfidf値のヒストグラムである．表3.3は，各色の下線文に含まれる語のtfidf値の中央値，四分位偏差値および語数を示している．まず初めに，Fligner-Killeen検定で3群の分散の差を調べた．その結果，$p < 0.05$となり，群間に有意差があることが分かった．赤・青・緑色の下線文に含まれる語のtfidf値群の中央値に差があるかどうかを，クラスカル・ウォリス検定で調べた．その結果，$p > 0.05$となり，群間の有意差を得ることはできなかった．したがって，ユーザは色を使い分けるときに，語の特徴度とは関係なく下線を付与していたことがわかった．

表 3.3: 各色の下線文に含まれる語に関するtfidf値

	赤線	青線	緑線
中央値	4.6	5.0	4.6
四分位偏差値	3.5	3.7	2.4
語数	141	174	141

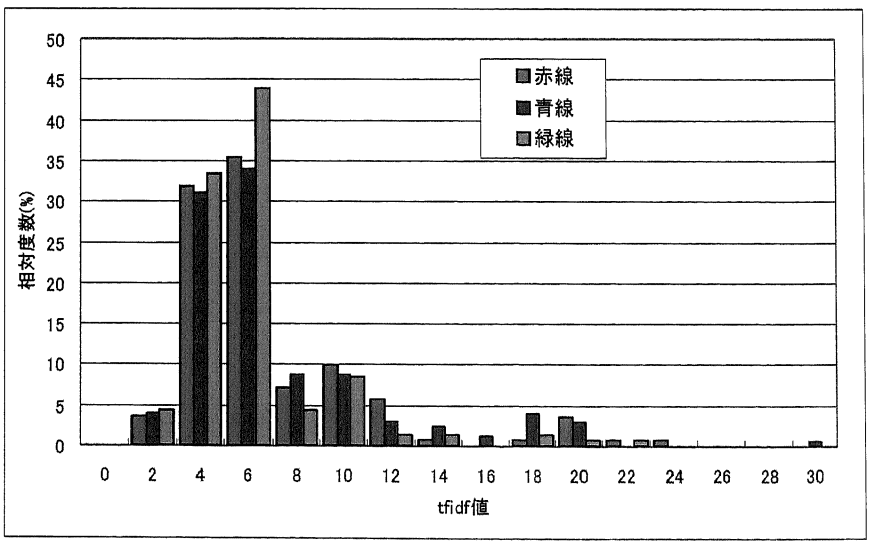

図 3.7: 各色の下線文に含まれる語に対するtfidf値のヒストグラム

3.3.3 各ユーザの下線の付け方

ここでは，各ユーザの下線の付け方について調べた．図 3.8 は，ユーザが下線を引いた数を示しており，各色の線の数を積み上げグラフにしたものである．27 人中，線を引いた数が多かった上位 7 人を載せている．

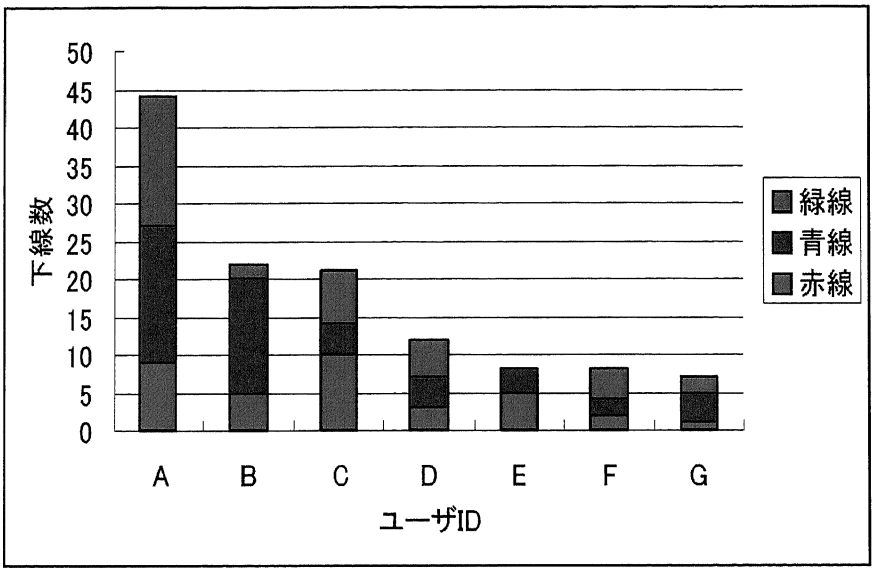

図 3.8: ユーザが引いた下線の数

次に，各ユーザが付与した下線に含まれる語に対する tfidf 値のヒストグラムを利用して，各ユーザがどのような語に下線を付与したいたのかについて述べる．

図 3.9 によると，ユーザ A は，青線と緑線を多く使っていた．したがってユーザ A は，客観的，あるいは主観的に重要な箇所の両方に下線を付与していたことが分かる．また，tfidf 値の一番高い値から低い値まで，広範囲に渡って各色の線を選んでいるので，ユーザ A が下線は語の特徴度とは関係なく下線を引いていたことを示している．

図 3.10 によると，ユーザ B は，赤色と青色の線ばかり選んでいる．データ区間が 18 から 20 の間の語の出現頻度が他のユーザよりも高いため，文書内の特徴語を客観的に重要だと判断したといえる．

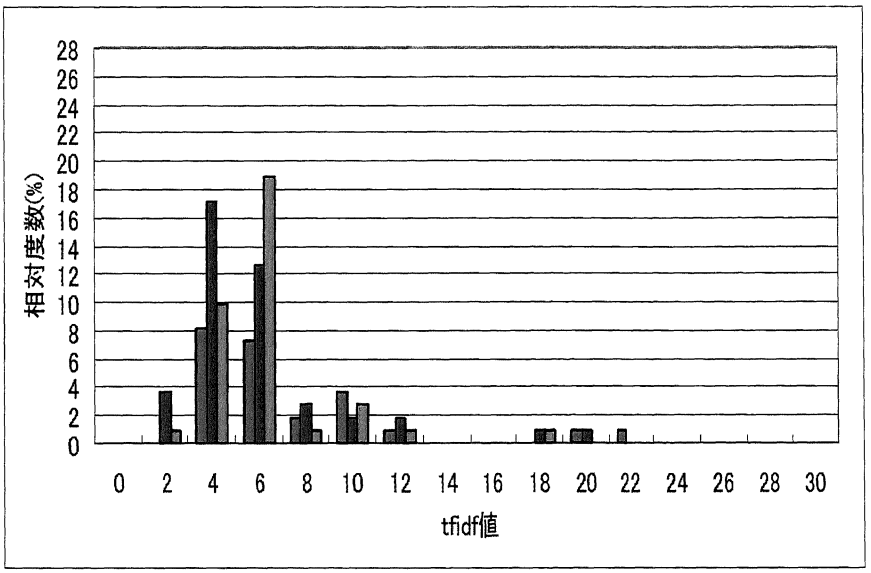

図 3.9: ユーザ A が引いた下線に含まれる語に対する tfidf 値のヒストグラム

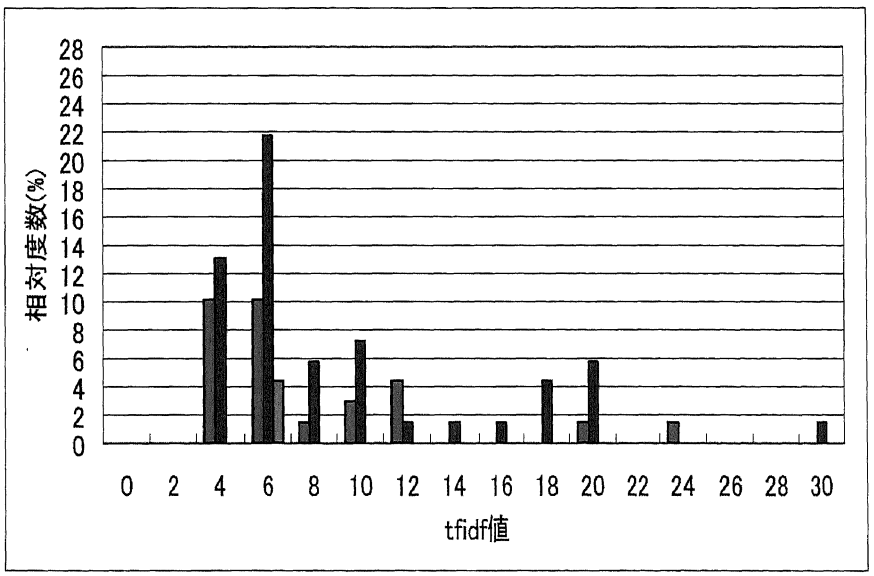

図 3.10: ユーザ B が引いた下線に含まれる語に対する tfidf 値のヒストグラム

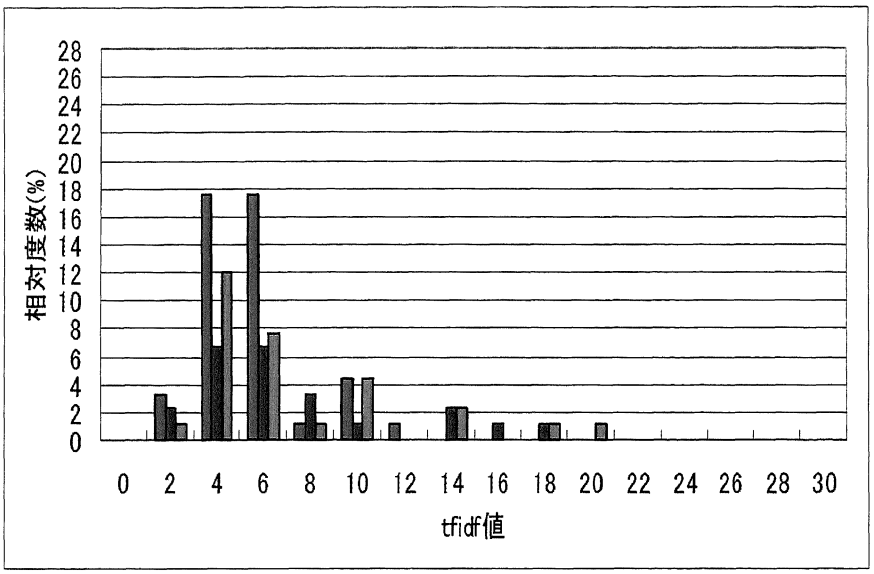

図 3.11: ユーザ C が引いた下線に含まれる語に対する tfidf 値のヒストグラム

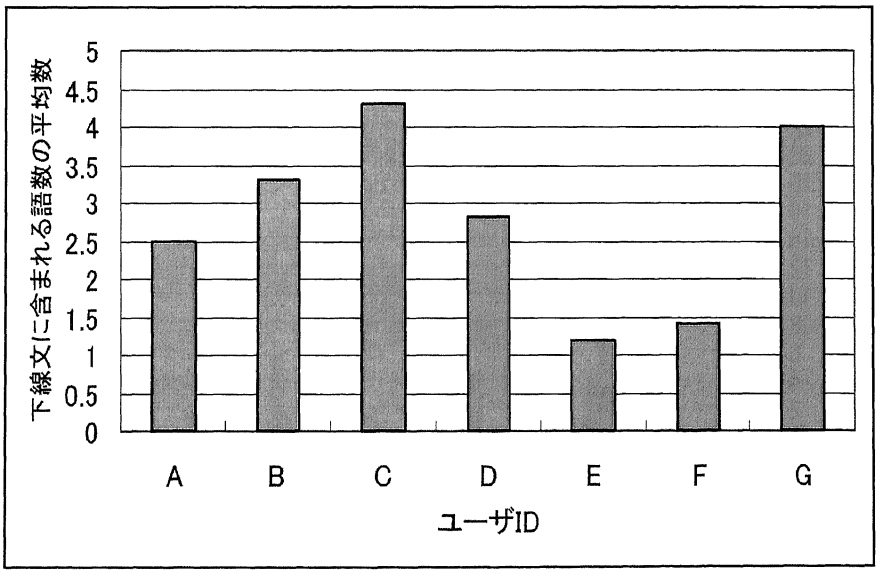

図 3.12: 下線文に含まれる語数の平均数

図 3.11 によると，ユーザ C は，赤色を一番多く選んでいる．他のユーザと比べると，

データ区間2から4において，語の出現頻度が若干高い．これは，ユーザCが長い下線を引く傾向があるからだと思われる．全ユーザの下線文に含まれる語数の平均数を示した図3.12によると，ユーザCは1番多いことが分かる．論文概要の内容を表す特徴的な語を1つ選ぶのではなく，一般語から構成される長い文章を好んで選んでいるため，データ区間が低いところの出現頻度が高い．一方で，tfidf値の高い区間において，緑線に含まれる語の出現頻度が高い．ユーザCは，文書内の特徴語を主観的に重要だと判断することが多かったことが分かる．

図3.13によると，ユーザDは，赤線よりも青と緑の下線を使用している．データ区間が4から6の間の語の出現頻度が飛び抜けて高いのが特徴である．これは，全論文概要に含まれる語のtfidf値の出現頻度が一番多いデータ区間と同じであり（図3.5），ユーザDは特徴度が中程度のものを好むことが分かる．また，tfidf値の高い区間において，緑線に含まれる語の出現頻度が高いことから，文書内の特徴語を主観的に重要だと判断する傾向があることが分かる．

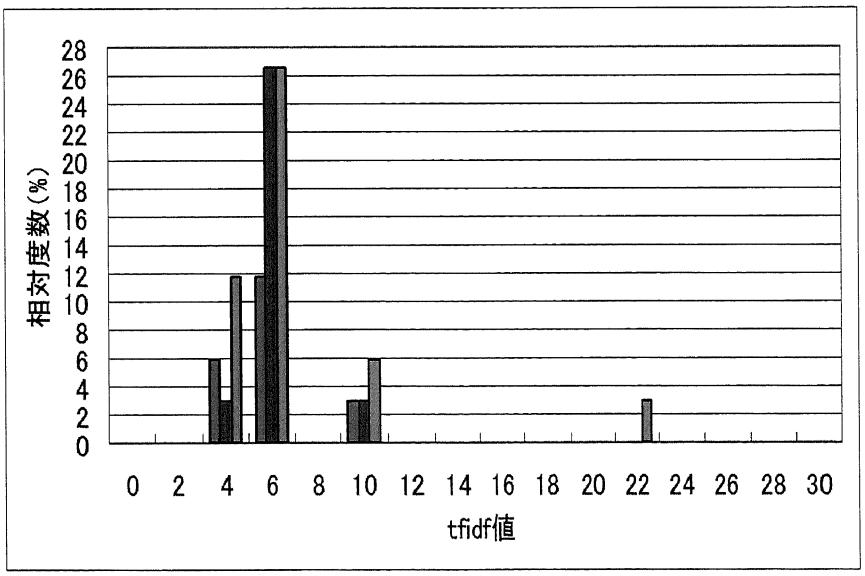

図3.13: ユーザDが引いた下線に含まれる語に対するtfidf値のヒストグラム

図3.14によると，ユーザEは，赤線と青線のみを使っていた．下線の数が少ないものの，データ区間6以上の語の出現頻度が高いため，文書内の特徴語を客観的に重要な語ととらえていることが分かる．図3.15によると，ユーザFに関しても，赤線に含まれる語

がデータ区間の高いところに多く出現する．したがって，ユーザ F も文書内の特徴語を客観的に重要な語ととらえていると考えることができる．

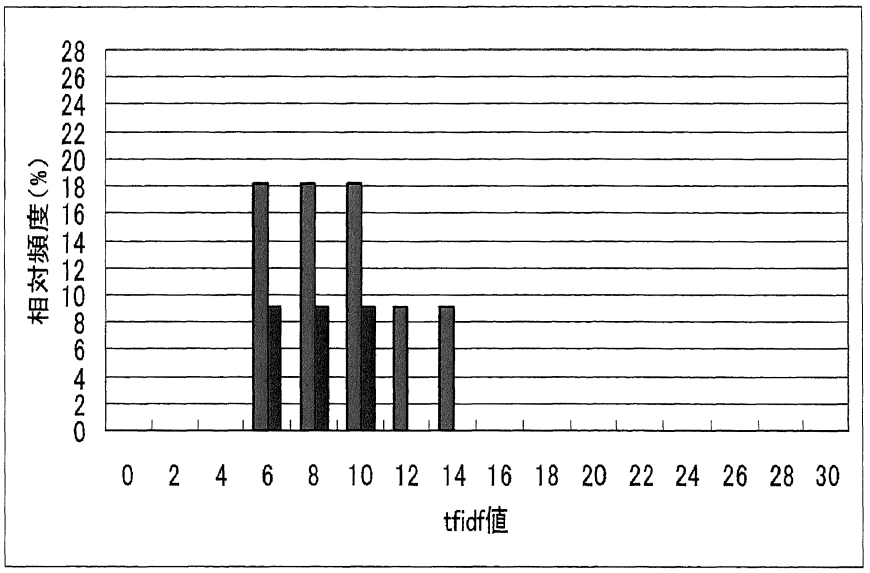

図 3.14: ユーザ E が引いた下線に含まれる語に対する tfidf 値のヒストグラム

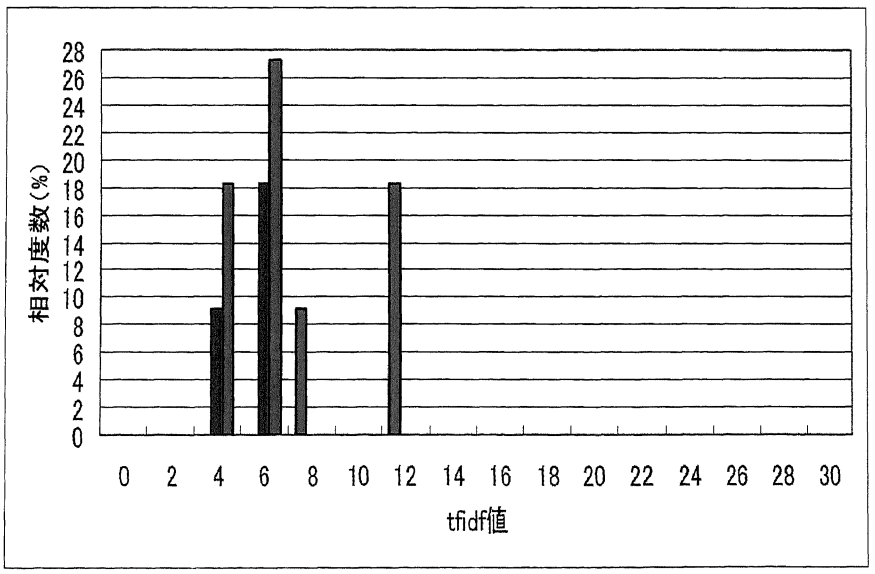

図 3.15: ユーザ F が引いた下線に含まれる語に対する tfidf 値のヒストグラム

図 3.16 によると，ユーザ G は，データ区間 4 における語の出現頻度が最も高い．ユーザ G は下線に含まれる語数が多いため，データ区間が低いところの語の出現頻度が高いものと思われる．また，赤線よりも緑線に含まれる語のほうがデータ区間の高いところに多く出現している．

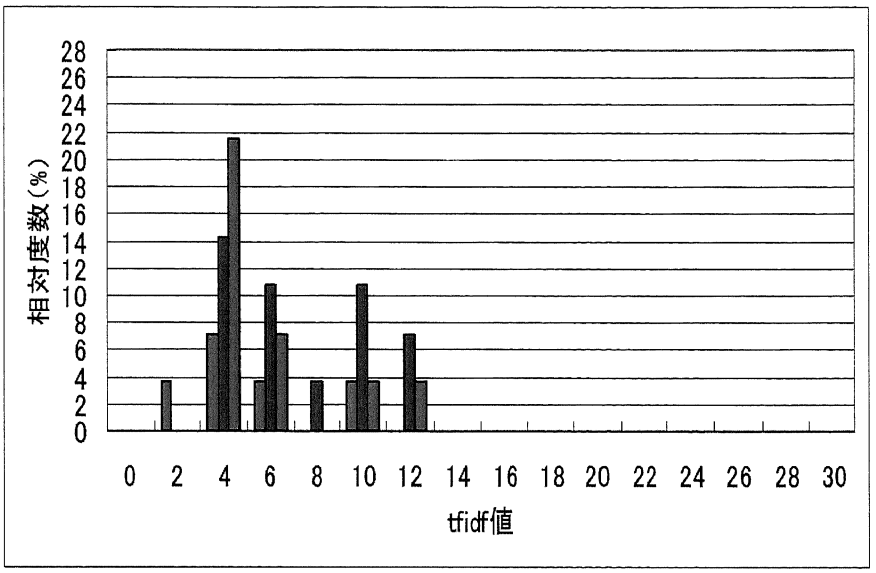

図 3.16: ユーザ G が引いた下線に含まれる語に対する tfidf 値のヒストグラム

これらユーザの下線の引き方をみると，長い下線文には tfidf 値の低い語が含まれやすいということがわかった．また，tfidf 値の高い語に赤色および青色の下線を付与するユーザもいれば，緑色の下線を付与するユーザもおり，下線の付け方はユーザによって様々であることがわかった．

3.3.4　下線が付与された語

本研究では，Web コンテンツの内容を直接反映した語を獲得するために，下線引きに着目した．1.3.4 で紹介したソーシャルブックマークタグの分類のうち，(1)〜(3) のタグが Web コンテンツの内容と直接関係あるタグであり，本研究で獲得したい語である．そこで本節では，ユーザが下線を付与した語が，(1)〜(3) の分類に該当するのかについて調査した．

ソーシャルブックマークタグの分類のうち (4)〜(7) は，個人的な意見や解釈が反映されたタグに関する分類であった．下線を付与することというのは著者が書いた語を選択することなので，下線を付与した語に個人的な意見を反映することはできない．従って，下線が付与された語は (4)〜(7) の分類には当てはまらない．

次に，下線が付与された語は，Web ページの内容と直接関係するタグである (1)〜(3) の分類に該当するのかについて述べる．今回は学会で発表された論文の概要文を対象としたので，下線が付与された語の中に (3) の Web ページを作成した人の名前にあたる語はなかった．しかし，著者名が書かれている文書，例えば，論文そのものやプロフィールが書かれた Web ページを対象にした場合は，人名に下線が付与される可能性がある．(2) の Web ページに書かれている内容の種類に関する分類においても，対象文書に「論文」という語は書かれていなかったため，該当する語はなかった．この場合，論文や Web ページに文書の属性が書かれていれば，下線が付与される可能性がある．

最後に，下線が付与された語は (1) の Web ページの主題に関する語に該当するのかについて述べる．下線が付与された回数が多い上位 10 個の語は，順に，「情報」，「ネットワーク」，「知識」，「状況」，「研究」，「行動」，「提示」，「クラスタリング」，「ユーザ」，「パターン」だった．これらの語は論文概要文に含まれる語なので，論文の内容と関係のある語，とは言えるが，主題に関する語であるとは言い切ることはできない．そこで，これらの語が論文の主題を表すタイトルとして使用されているかどうかを調べた．論文のタイトルに含まれる回数が多い上位 30 個の名詞のうち，「情報」，「ネットワーク」，「知識」，「行動」，「クラスタリング」，「ユーザ」の 6 語が含まれていた．このように，ユーザが下線を付与した語のなかには主題語も含まれることから，下線が付与された語は（1）の Web ページの主題に関する語に該当する場合があるといえる．これらの考察より，下線が付与された語は (1)〜(3) の分類に該当する可能性があることから，Web コンテンツの内容と直接関係する語として利用できるものと思われる．

3.4 考察

3.3.1 から 3.3.3 にかけて，文章内の語の特徴度を示す指標である tfidf を用いて，下線が付与された語について各種分析を行った．3.3.1 における分析結果より，全ユーザが付与した下線文には，論文概要文よりも tfidf 値の高い語が多く含まれることが分かった．3.3.2 では，ユーザが付与した色と語の特徴度とに関連性があるのかについて調べたところ，各色線に含まれる語の特徴度に有意差はなかった．3.3.3 では，各ユーザの下線の付け方につ

いて調べたところ，tfidf値の高い語に赤色および青色の下線を付与するユーザもいれば，緑色の下線を付与するユーザもおり，下線の付け方はユーザによって様々であることが分かった．

3.3.1より，全ユーザが付与した下線文を集約すると，tfidf値の高い語が多く含まれることから，文書内における特徴度の高い語が下線引きによって選択されていたといえる．本研究では，ユーザが情報を探しやすくするために，コンテンツ内の特徴語が記述された意味的なメタデータを生成することを目標としているため，下線文に含まれる語を意味的なメタデータとして利用できる可能性があることが分かった．また，tfidfは文書集合の数が分かっているときに算出できる指標なので，どれくらいの数のWebコンテンツが存在するか分からないWWWにおいて利用できる指標ではない，しかし，下線が付与された文にはtfidf値の高い語が多く含まれていたことから，下線が付与された語をtfidf値の高い語とみなすことによって，tfidfの代用として下線引きを利用できるかもしれない．

3.3.2において，各色線に含まれる語の特徴度に有意差がなかったのは，3.3.3で示したように，ユーザによって下線の使い方が異なっていたからだと思われる．齋藤は著書で「個人的な観点と一般的な観点とがもちろん重なる場合がある．そのときには，緑と赤が同一箇所に引かれることになる．二色が重なるのはまったく構わない.」と述べている．イロノミーでは同一箇所に複数の色線を付与することができないため，ユーザによっては客観的に重要な箇所を主観的に重要な箇所として色線を付与していた可能性もある．システムの運用で得られたデータからは，ユーザがどういう意図で色を使い分けていたのかまでは正確に推測することはできない．したがって，分析結果からいえることは，各ユーザによって色の付け方は異なるが，ユーザ全体でみると平均化されて各色線に含まれる語のtfidf値の有意差がなくなる，ということである．この結果より，イロノミーでは集合知が機能していたといえる．

3.3.4では，下線が付与された語がコンテンツの内容と直接関係ある語なのかどうかについて調査した．ソーシャルブックマークタグの分類を利用して調査したところ，下線が付与された語はコンテンツの内容と直接関係ある語の可能性が高いことが分かった．3.3.3より，長い下線文にはtfidf値の低い語，すなわち一般語が含まれやすいことが分かっている．したがって，下線が付与された語すべてが，タグの分類の(1)Webページの主題に関する語に該当するとは限らない．下線が付与された文字列が長い場合，メタデータに記述する語としてどの語に注目すべきなのかを判断する必要がある．

3.5 まとめ

本章では，複数ユーザによって下線が付与された語を意味的なメタデータとして利用できるかについて調査した．調査には，三色ボールペン読書法に基づいて色線を付与できるシステム，イロノミーを利用した．イロノミーの運用で得られたデータの分析より，全ユーザで見ると色にかかわらずtfidf値の高いことから，特徴語に下線が付与される可能性が高いということが分かった．また，ソーシャルブックマークタグの分類を利用した考察によると，下線が付与された語はWebコンテンツの内容と直接関係ある語であることと，主題を表す語の場合があることが分かった．これらの結果より，複数ユーザが付与した下線文を集約すると，コンテンツ内の特徴語が多く，主題を表す語が含まれることから，意味的メタデータの生成に有用な語を獲得できる可能性があることが見出された．

第 4 章

マーキングの共有による情報探索の有効性に関する分析

本章では，ユーザが下線を付与した語，すなわち，マーキングを付与した語や文字列を他人と共有した際，情報探索に役立つのかについて調べた．調査に利用したのは，第 20 回人工知能学会全国大会 (JSAI2006) で運用されたシステム “合口” である．本章では，合口の運用で得られたデータを使って行った分析について述べる．

4.1 はじめに

ソーシャルブックマークでは，複数ユーザが付与したタグをユーザ間で共有することで，タグを通じて情報を探すことができた．3 章より，複数ユーザが Web コンテンツに付与した下線文を集約すると，コンテンツの特徴語が多く含まれるということがわかったことから，マーキングをユーザ間で共有した場合，情報探索に役立つ可能性がある．そこで本章では，Web コンテンツに付与されたマーキングを複数ユーザ間で共有したときに，ユーザがマーキングによってコンテンツ内で選択した語や文字列がソーシャルタギングの役割を果たすのかどうかについて調査した．

4.1.1 関連研究

ユーザが付与したアノテーションが他人に役立つかどうかに関する調査はいくつかなされており，それによるとその分野のエキスパートが付与したアノテーションは読者による再現率が高いという結果が得られている．これは他人によるアノテーションがコンテンツの理解に役立つかの調査であり，情報探索にマーキングを利用する場合，エキスパートによる情報だけが必ずしも役立つのかといった疑問がある．そこで本研究では，人工知能学

会全国大会において，先生だけでなく学生もいるという状況を利用して，他人のアノテーションが情報探索をする際に活用できるのかについて調査した．

また，Web コンテンツ内の文字列をマウスカーソルで選択することによってアノテーションを付与できるシステム，YAWAS における調査によると [Denoue 00]，レポートの課題がある学生たちが，アノテーションが付与された文字列に対する検索エンジンを利用する際，検索クエリを空白にすることによって得られたアノテーションが付与された文字列をコンテンツのサマリーとして利用していたという結果が報告されている．この場合，マーキングを付与した文字列のみを検索対象にすると，検索クエリが検索対象に含まれない場合は結果が返ってこないという問題がある．したがって，検索結果がいつでも表示されるような工夫が必要である．

4.2 合口

本研究では，他人が付与したマーキングが情報探索に役立つのかについて調査するために “合口” というシステムを提案した．合口は，ユーザ発表ページ内の文字列をマーキングすると，他の発表ページを推薦するシステム [松岡 07] である．

合口の操作方法について述べる．合口では，ユーザが発表ページ内の文字列をマウスカーソルで選択した際 (図 4.1)，推薦アルゴリズムに基づいて算出された他の発表ページへの推薦リンク（発表ページのタイトル）が書かれた小窓を表示する (図 4.2)．ユーザは提示された推薦リンクの中から気に入ったものをクリックをすると，クリック先の発表ページへ遷移すると同時に，合口はユーザが選択した文字列をマーキング文字列として発表ページ上に付与する (図 4.3)．発表ページに付与されたマーキング文字列は，ハイライト表示されて他の発表ページへのリンクアンカの役割を果たす．このマーキング文字列をユーザがマウスカーソルでなぞると，合口は足跡リンクと推薦リンクが書かれた小窓を表示する (図 4.4)．足跡リンクは，以前誰かがこのマーキング文字列 (選択文字列) から遷移したことのある他の発表ページへのリンクで，推薦リンクは推薦アルゴリズムに基づいてシステムが推薦した他の発表ページへのリンクである．

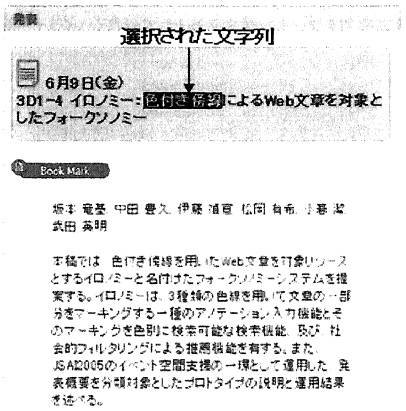

図 4.1: ユーザが Web ページ内の文字列を選択する

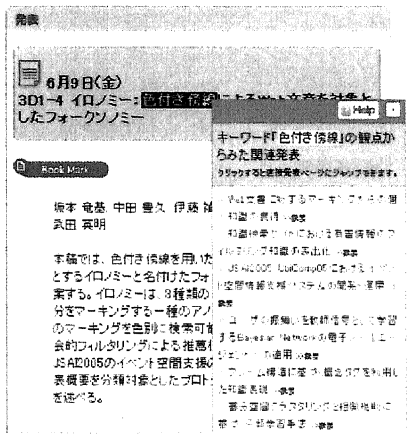

図 4.2: 推薦リンクが書かれた小窓を表示する

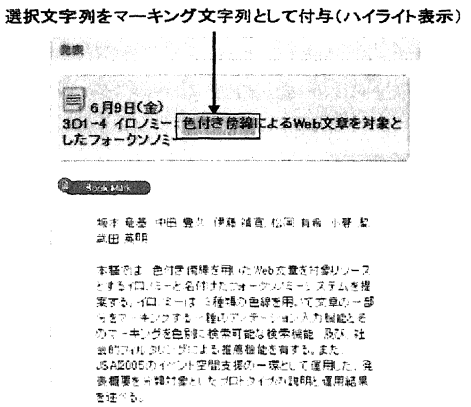

図 4.3: 選択文字列をマーキング文字列として発表ページ上に付与する

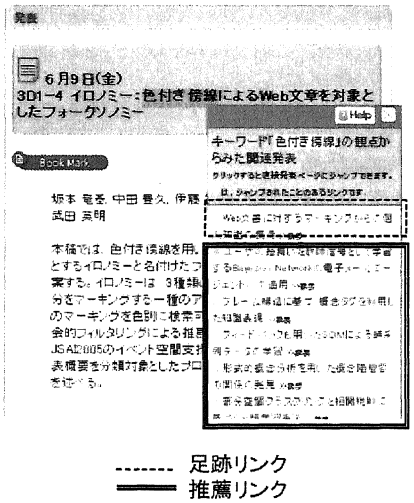

図 4.4: 足跡リンクと推薦リンクが書かれた小窓を表示する

ここでは，ユーザがマーキングを付与した文字列内の語をタグとみなした場合のソーシャルタギングとしての有効性を検証するために，4 種類の推薦アルゴリズムを使った他ページへの推薦機能を実装した (図 4.2)．また，ユーザがマーキングを付与した文字列をタグとみなした場合のソーシャルタギングとしての有効性を検証するために，足跡リンクと推薦リンクを同時に表示するよう実装した (図 4.4)．

4.2.1 推薦アルゴリズム

マーキングが付与された文字列内の語が情報探索に有用かどうかを調べるために，4種類の推薦アルゴリズムを用意した．用意した推薦アルゴリズムは下記のとおりである．

A) tfidfを使ったページ間類似度による推薦

B) 発表ページにユーザが付与したマーキング文字列の数を使った協調フィルタリングによる推薦

C) ユーザがマウスカーソルで選択した文字列内の語と他の発表ページに付与されているマーキング文字列内の語とのマッチングによる推薦

D) ユーザがマウスカーソルで選択した文字列内の語と他の発表ページ内の語とのマッチングによる推薦

アルゴリズムAでは，ユーザがマウスカーソルで文字列を選択した発表ページに対して，tfidfを使ったページ間類似度が高い他の発表ページを推薦する．イロノミーの分析と同様に，各論文概要文ごとに茶筌を使って形態素解析をして得られた2文字以上の名詞および未知語を使ってtfidf値を求めた．tfidf値は，式3.1により求めた．論文概要ベクトルは，以下のようになる．

$$v_{a_i} = \{tfidf(w_1, a_i), tfidf(w_2, a_i), \ldots tfidf(w_j, a_i)\}$$
$$(1 \leq i \leq N, 1 \leq j \leq WN) \quad (4.1)$$

aiは，発表ページを表しており，発表ページ数Nは276ページである．wjは，発表ページ内に含まれる語を表しており，WNは発表ページに含まれる語数である．類似度は2つの論文概要ベクトルの内積によって求める．類似度は，以下の式により求める．

$$simi(v_{an}, v_{am}) = \frac{v_{an} \cdot v_{am}}{\|v_{an}\| \|v_{am}\|} \quad (4.2)$$

アルゴリズムAは，ユーザが選択した文字列や他のページに付与されているマーキング文字列とは関係なく，ユーザがマウスカーソルで文字列を選択した発表ページに対して，類似度の高い発表ページを推薦する．

アルゴリズムBでは，協調フィルタリング[Resnick 94]を用いて，似た嗜好を持ったユーザが選択したことのある発表ページを推薦する．ユーザがマーキング文字列を付与した発

表ページのうち，同じ発表ページにマーキングを付与したことのあるユーザ同士は似た嗜好を持っている可能性が高い．そこで，ユーザによる発表ページへの評価値を，発表ページ上にユーザが付与したマーキング文字列の数とし，協調フィルタリングにより他の発表ページを推薦する．計算は，Nearest neighbor 法に基づいた．手順は下記のとおりである．まず初めに，ユーザ間の類似度を，評価値ベクトル（Web ページ内に付与されたマーキングの数）間の Pearson 相関係数として計算する．ユーザ $u1$ と $u2$ との間の類似度 $S_{u1,u2}$ は以下のように求める．

$$S_{u1,u2} = \frac{\sum_{i=1}^{N} (r_{u1,i} - \bar{r}_{u1}) \times (r_{u2,i} - \bar{r}_{u2})}{\sqrt{\sum_{i=1}^{N} (r_{u1,i} - \bar{r}_{u1})^2 \times \sum_{i=1}^{N} (r_{u2,i} - \bar{r}_{u2})^2}} \tag{4.3}$$

ここで，$r_{u1,i}$ はユーザ $u1$ によって項目 i に与えられた評価値であり，$\bar{r}_{u1}$ はユーザ $u1$ によって与えられた評価値の平均である．また，N は項目の総数を表しており，ここで発表ページの総個数を表している．次に，Web ページ内の文字列を選択したユーザに関して，最も高い類似度をもつ n 人のユーザを近傍のユーザとして選択し，近傍のユーザの評価値の重み付けされた組み合わせから予測値を以下のように計算する．

$$P_{u1,i} = \bar{r}_{u1} + \frac{\sum_{u=1}^{n} (r_{u2,i} - \bar{r}_{u2}) \times S_{u1,u2}}{\sum_{u=1}^{n} S_{u1,u2}} \tag{4.4}$$

ここで，$P_{u1,i}$ は対象とするユーザ $u1$ の項目 i に対する予測値を，$S_{u1,u2}$ はユーザ $u1$ と $u2$ との間の類似度を，n はユーザ $u1$ の近傍におけるユーザ数を表す．アルゴリズム B は，発表ページに付与されているマーキング文字列の数を用いるが，マーキング文字列内の語は一切考慮しない．

アルゴリズム C では，ユーザが発表ページ内で選択した文字列内の語と他の発表ページに付与されているマーキング文字列内の語とのマッチングを行って，マッチした場合に発表ページを推薦する．他の発表ページ上に付与されているマーキング文字列は全ユーザによって付与されたものを対象とする．また，マッチングに利用する語は，選択文字列内およびマーキング文字列内の名詞および未知語である．

アルゴリズム D では，ユーザが発表ページ内で選択した文字列内の語と他の発表ページ内の語とのマッチングを行ってマッチした発表ページを推薦する．アルゴリズム D は，ユーザが発表ページ内で選択した文字列内の語を検索クエリとし，他の発表ページ内に含まれているかどうかを調べている．一般的にユーザが Web ページを探すのに最も利用するのは検索エンジンであるため，検索エンジンで行われることと同じ手法を推薦に取り入れた．

各推薦アルゴリズムにおいて，選択文字列（ユーザがマウスカーソルで選択した文字列）や発表ページに付与されたマーキング文字列，発表ページ内の文字列を使用するかどうかを表 4.1 にまとめた．

表 4.1: 各推薦アルゴリズムおいて使用する文字列の比較

推薦アルゴリズム	選択文字列	マーキング文字列	発表ページ内の文字列
A	×	×	○
B	×	○	×
C	○	○	×
D	○	×	○

アルゴリズム A のページ間の類似度はシステムの運用前にあらかじめ計算しておき，その他の推薦アルゴリズムに関しては合口の運用中に動的に計算した．合口はユーザが発表ページ内の文字列をマウスカーソルで選択すると，各アルゴリズムにつき最大 2 つのページを推薦し，表示はランダムに並べた．同じ発表ページが異なる推薦アルゴリズムによって算出されたときのために，推薦アルゴリズムに優先度を設けた．優先順位は，C，D，B，A の順である．ユーザにはこれらの推薦アルゴリズムや表示方法については知らせていない．ユーザがどの推薦アルゴリズムを選択したのかについては，ユーザが合口によって推薦された他の発表ページへのリンクをクリックした時に，そのリンクを推薦するために用いたアルゴリズムを選択したとする．

4.2.2 運用結果

合口が対象とした Web ページは論文のタイトルや発表者，概要を含む発表ページで，全部で 276 ページあった．合口は学会の開催前から運用しており，分析対象としたデータは，2006 年 5 月 22 日 (月)〜6 月 9 日 (金) までの運用によって得られたデータである．運用の結果，開発者を除く 40 人のユーザが 1 回は発表ページ内の文字列をマウスカーソルで選択し，そのうち 27 人が提示された推薦リンクをクリックした．また，開発者を除く 83 人のユーザが 1 回は発表ページ上のマーキング文字列をマウスカーソルでなぞり，そのうち 32 人が提示されたリンクをクリックした．

4.3 分析

本節では，マーキングが付与された文字列を複数ユーザ間で共有したときに，情報探索に有益かどうかについて分析を行った．また，3 章より，ユーザが下線を付与した文字列にはtfidf値の高い語が多く含まれることがわかったので，実際にtfidf値の高い語が情報探索に有用なのかについても調べた．

4.3.1 マーキングが付与された文字列内の語が情報探索に有益かどうか

ここでは，マーキングが付与された文字列を複数ユーザ間で共有した場合，マーキングが付与された文字列内の語が情報探索に有益かどうかを調べた．具体的には，ユーザが発表ページ内の文字列をマウスカーソルで選択したときに，システムが推薦した他の発表ページのうち，どの推薦アルゴリズムによる推薦を選択したのかについて調査した．ユーザがマウスカーソルで発表ページ内の文字列を選択したときにシステムが提示した推薦リンクをクリックしたことがあるユーザのうち，学会前に使用していたのは 20 人で，学会前だけ使用していたユーザは 14 人だった．一方で，学会中に使用していたのは 13 人で，学会中だけ使用していたユーザは 7 人だった．このように，学会前と学会中とでシステムの利用者が異なるため，2 つの期間に分けて調査をした．

表 4.2 は，ユーザが発表ページ内の文字列をマウスカーソルで選択したときに，システムが各推薦アルゴリズムによって推薦した発表ページの数と，推薦された発表ページのうちユーザが選択した発表ページの数を示している．また，図 4.5 は，各推薦アルゴリズムによって推薦された発表ページのうちユーザが選択した割合 (ユーザが選択した発表ページ数/システムが推薦した発表ページ数× 100) を示している．

表 4.2: システムが各推薦アルゴリズムによって推薦したページ数とユーザによって選択されたページ数

推薦アルゴリズム	A	B	C	D
学会前 (5/22-6/6)	27/307	9/129	7/118	10/238
学会中 (6/7-6/9)	7/117	1/47	11/66	11/103

表の値は，ユーザが選択したページ数/システムが推薦したページ数

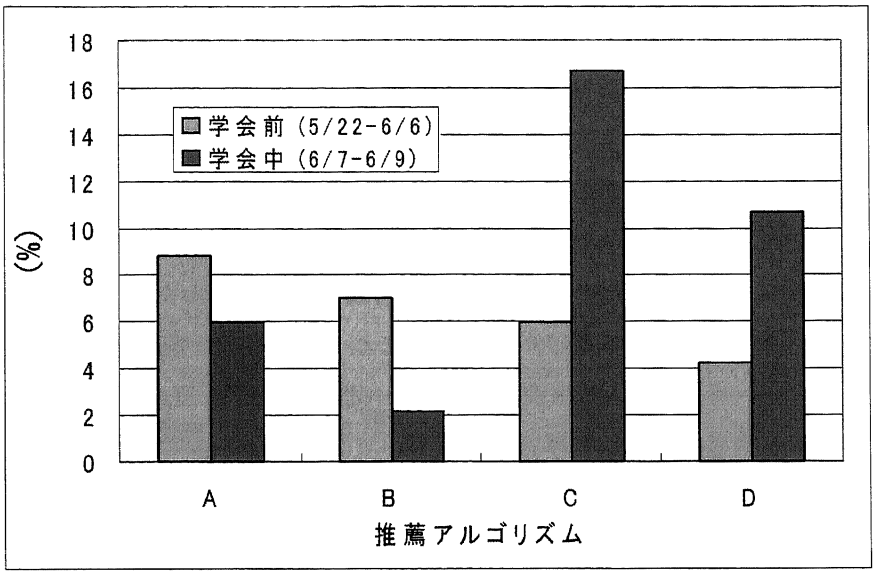

図 4.5: 各推薦アルゴリズムによって推薦された発表ページのうちユーザが選択した割合

これによると，ユーザが学会前に最も選択した推薦アルゴリズムはＡで，次は推薦アルゴリズムＢである．学会前にシステムを使用したユーザは，マウスカーソルで選択した文字列内の語や他の発表ページに付与されているマーキング文字列内の語を使わない推薦による発表ページを選択していた．

一方で学会中にシステムを使用したユーザは，推薦アルゴリズムＣやＤによる推薦が推薦アルゴリズムＡによる推薦よりも選択していたことから，ページ間類似度の高い発表ページよりもユーザが発表ページ内で選択した文字列内の語が含まれる他の発表ページを好んだといえる．なかでも推薦アルゴリズムＣによって推薦されたリンクがもっとも選択されていたことから，学会中にシステムを利用したユーザは，マーキング文字列内の語を利用したWebページを好むということが分かった．

学会前にシステムを利用したユーザは，メールによるシステム運用の告知があった次の日に，システムに推薦されたページを多く選択していたので[*1]，試しにシステムを利用したユーザが多かったものと思われる．学会前に選択された各推薦アルゴリズムを見ても，選択された割合に目立った差はないため，ユーザは推薦されたページの中からランダムに選択した可能性がある．

[*1] 学会前に推薦ページが選択された回数の一日の平均値は 4 回で，システム運用の告知があった次の日に推薦ページが選択された回数は 15 回だった．

一方で，学会中にシステムを利用したユーザは，推薦アルゴリズムＣとＤによる推薦ページを選ぶ割合が他の推薦アルゴリズムによるものより高いため，マウスカーソルで選択した文字列内の語と関連のあるページを選択することが示唆された．これは，学会中にシステムを利用したユーザは文書内で注目した語に関するページを探すという目的を持って使用したものと思われる．このような状況においては，マーキングが付与された語が情報探索に有益である可能性がある．

次に，推薦リンクの表示順位がユーザの選択行為に影響があったのかについて調べた．合口では，ユーザに対し，各推薦アルゴリズムによって算出した推薦リンクをランダムに表示していた．ユーザは単に上位に表示されていた推薦リンクを選んでいたかもしれない．そこで，ユーザが推薦リンクを選ぶときに，表示順位に影響されていたのかについて調べた．図 4.6 は会期前における各推薦アルゴリズムによって算出された推薦リンクをユーザが選択したときの表示順位と推薦回数を示しており，図 4.7 は選択回数を示している．図 4.8 は会期前における各推薦アルゴリズムによって算出された推薦リンクをユーザが選択したときの表示順位と推薦回数を示しており，図 4.9 は選択回数を示している．

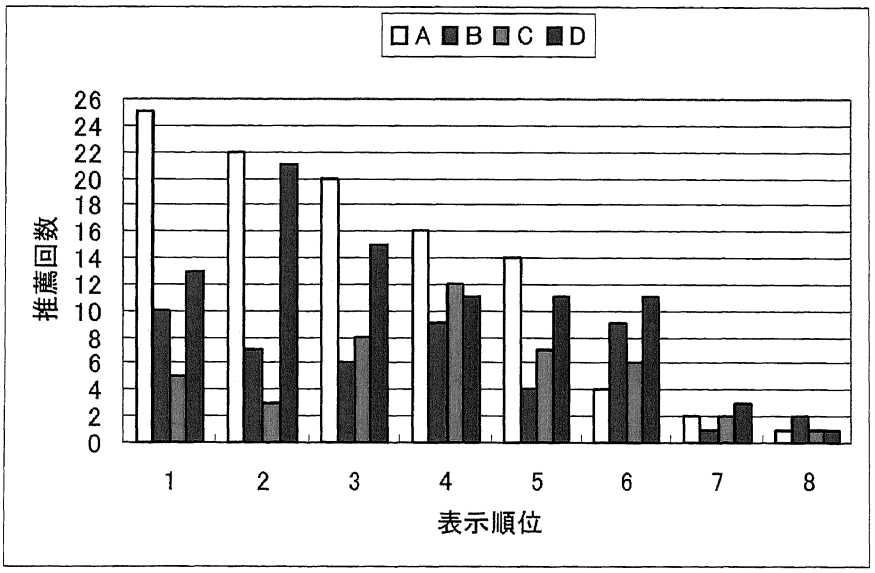

図 4.6: 会期前における各推薦アルゴリズムによって算出された推薦リンクをユーザが選択したときの表示順位と推薦回数

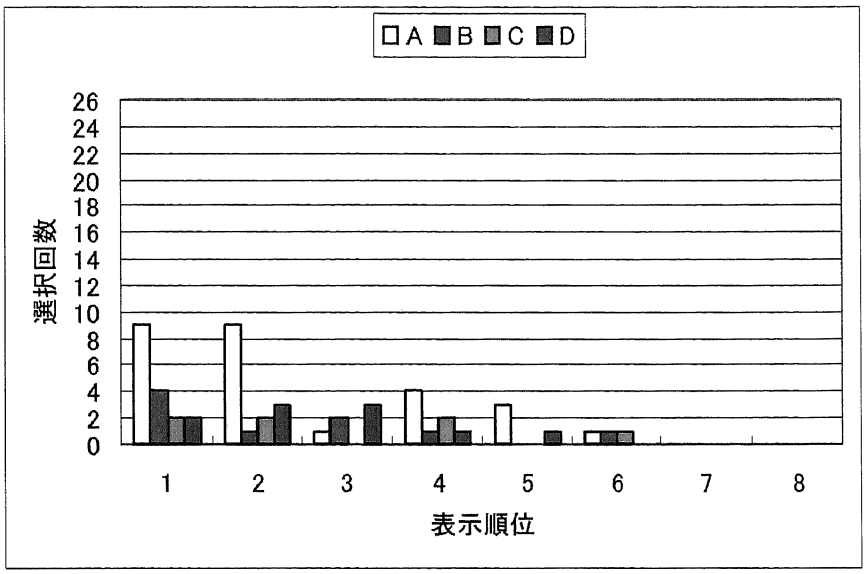

図 4.7: 会期前における各推薦アルゴリズムによって算出された推薦リンクをユーザが選択したときの表示順位と選択回数

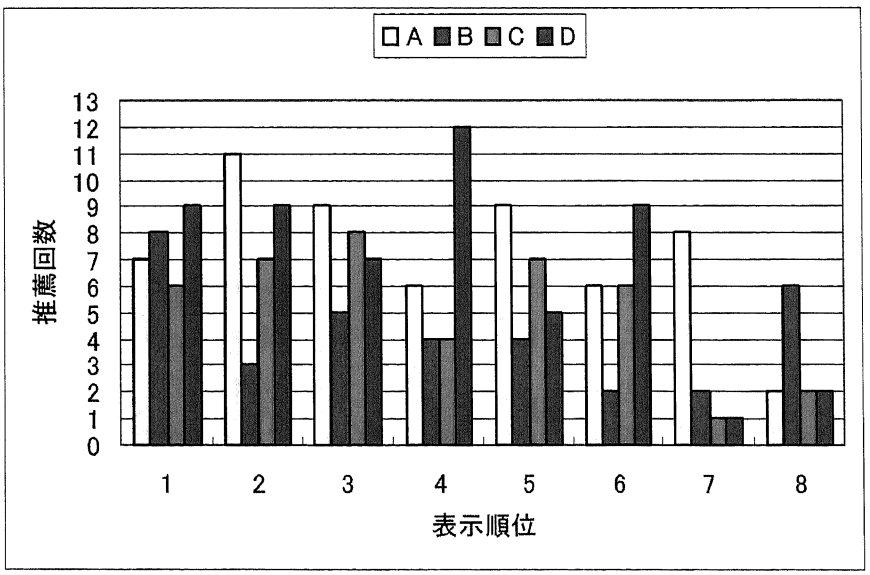

図 4.8: 会期中における各推薦アルゴリズムによって算出された推薦リンクをユーザが選択したときの表示順位と推薦回数

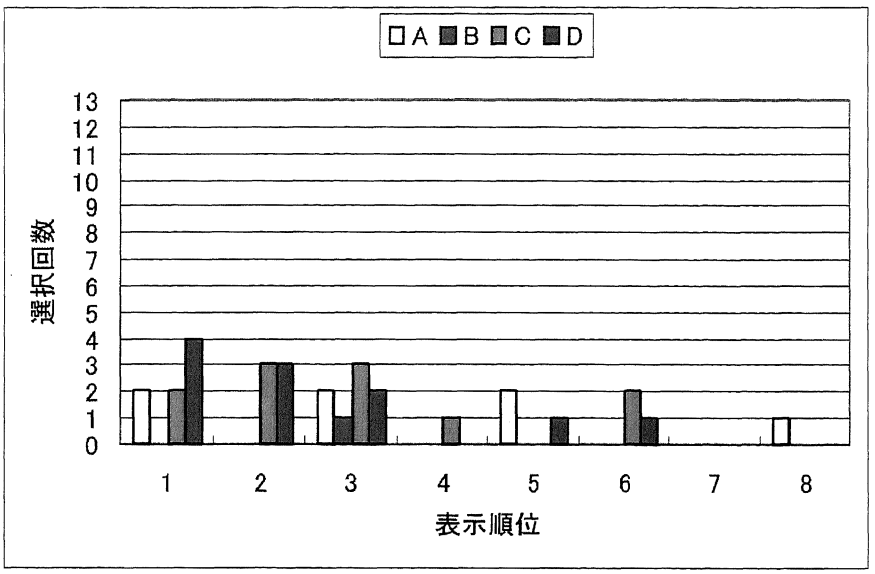

図 4.9: 会期中における各推薦アルゴリズムによって算出された推薦リンクをユーザが選択したときの表示順位と選択回数

図 4.7 のによると，会期前は表示順位 1，2 において，推薦アルゴリズム A による推薦リンクが多く選ばれている．これは，システムの運用の初期段階において，推薦アルゴリズム A による推薦リンクが推薦される回数が多かったこと（表 4.2）が，上位で推薦されたことと関係あるものと考えられる．推薦アルゴリズム B は，一番目に表示された推薦リンクの選択回数が多いが，下位で表示された推薦リンクが選択された回数とそれほど変わらない．推薦アルゴリズム C は，1，2，4 番目に表示された推薦リンクの選択回数が同じである．推薦アルゴリズム D は，1 番目よりも 2，3 番目に表示されたときのほうが選択回数が多い．これらの考察より，会期前において，表示順位がユーザの推薦リンクの選択行為に影響を与えたとは考えにくい．図 4.7 より，会期中においても，一番目に表示された推薦リンクの選択回数が多いのは推薦アルゴリズム D によるものだけである．したがって，ユーザが表示された順位にかかわらず，推薦されたリンクの文字列を見て選択していたものと思われる．

最後に，ユーザが選択した文字列内の語が推薦リンクの文字列に含まれているかどうかについて調べた．合口では，ユーザに推薦リンクを提示する際，発表論文のタイトルの文字列を提示していた．そこで，ユーザは選択した文字列内の語が推薦リンクの文字列に含

まれるかどうかで，推薦リンクを選んでいたのかどうかについて調べた．図 4.10 は，ユーザが選択した文字列に推薦リンクの文字列が含まれている割合を表している．

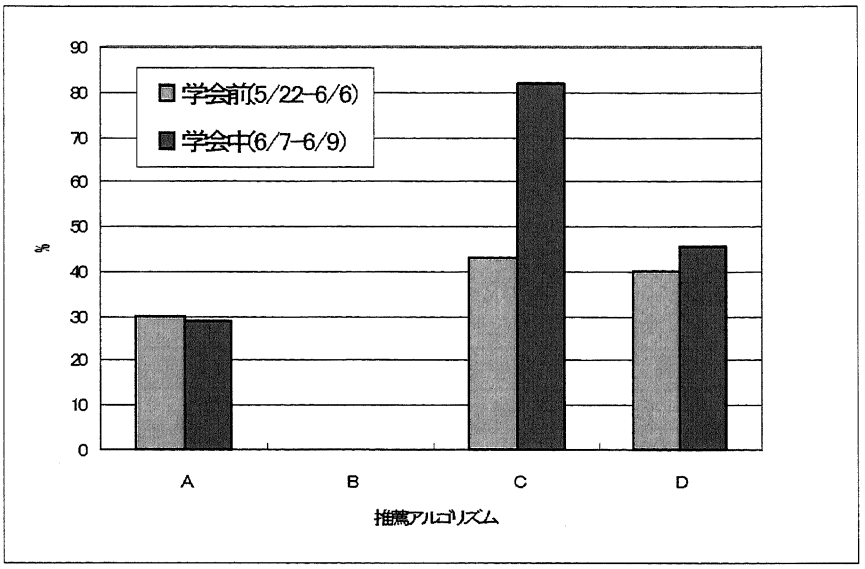

図 4.10: 選択文字列に推薦リンクの文字列が含まれている割合

これによると，会期中の推薦アルゴリズム C において，推薦リンクの文字列に選択文字列内の語が含まれる割合が最も高い．したがってこの結果は，前述した，ユーザは文書内で注目した語に関するページを探すという目的を持って使用した，ということを証明している．他の推薦アルゴリズムに関しては，50%以下の割合で選択文字列内の語が推薦リンクの文字列に含まれているので，推薦リンクの文字列に選択文字列内の語が含まれていることによってリンクが選ばれているとはいえない．

4.3.2 マーキングが付与された文字列が情報探索に有益かどうか

ここでは，マーキングが付与された文字列をユーザ間で共有した際，情報探索に有益かどうかについて調査した．合口では，ユーザが発表ページ上に付与されているマーキング文字列をマウスカーソルでなぞると，足跡リンクと推薦リンクが書かれた小窓を表示した (図 4.4)．足跡リンクはユーザがマーキング文字列から他の発表ページへ張ったリンクなので，マーキング文字列と直接関係のあるページとみなすことができる．そこで，ユーザが

マーキングを付与した文字列から他のページへ張られた足跡リンクと，システムが機械的に推薦をした推薦リンクのどちらが好まれたのかについて調べた．

表 4.3 は合口が推薦リンクや足跡リンクとして推薦したページ数と，推薦された発表ページのうちユーザが選択したページ数を示している．推薦リンクや足跡リンクとして推薦された発表ページのうちユーザが選択した割合を示した図 4.11 によると，学会前と学会中の両方の期間においてユーザは推薦リンクよりも足跡リンクによって提示されたページを選択していた．

表 4.3: システムが推薦リンクや足跡リンクとして推薦したページ数とユーザが選択したページ数

推薦アルゴリズム	推薦リンク	足跡リンク
学会前 (5/22-6/6)	20/2825	53/682
学会中 (6/7-6/9)	9/2468	17/622

表の値は，ユーザが選択したページ数/システムが推薦したページ数

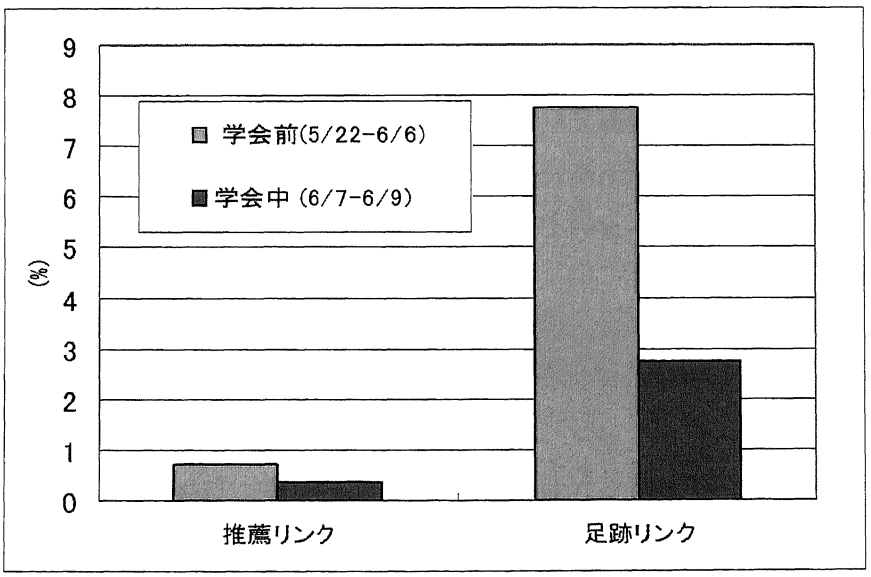

図 4.11: 推薦リンクや足跡リンクとして推薦された発表ページのうちユーザが選択した割合

合口では図 4.4 に示すように，足跡リンクの意味合いをユーザに明示化していたことか

ら，ユーザは意図して推薦リンクよりも足跡リンクを選んだ可能性がある．この場合，マーキング文字列がソーシャルタギングとして機能していたと言える．しかし，[Joachims 05]らは，ユーザは検索結果の表示一覧から何番目のリンク先へ遷移するかを調査した結果，一番上に表示されているクリックする傾向が極めて高いと結論づけている．合口では，足跡リンクを常にリストの最上段に表示したため，ユーザは単に上に表示されていた足跡リンクを選択した可能性も考えられる．したがってユーザがどういう意図で足跡リンクのほうを選択したのかは定かではないが，マーキングが付与された文字列が情報探索に有益であることが示唆された．

さらに，ユーザによって選択された足跡リンクを推薦するために用いられた推薦アルゴリズムついて調べた．ユーザによって選択された足跡リンクは，会期前で53個，会期中で17個あった．これらの足跡リンクがどの推薦アルゴリズムによって推薦されたのかについて調べた．図4.12は，ユーザが選択した足跡リンクを推薦するために用いられた推薦アルゴリズムの割合を示している．

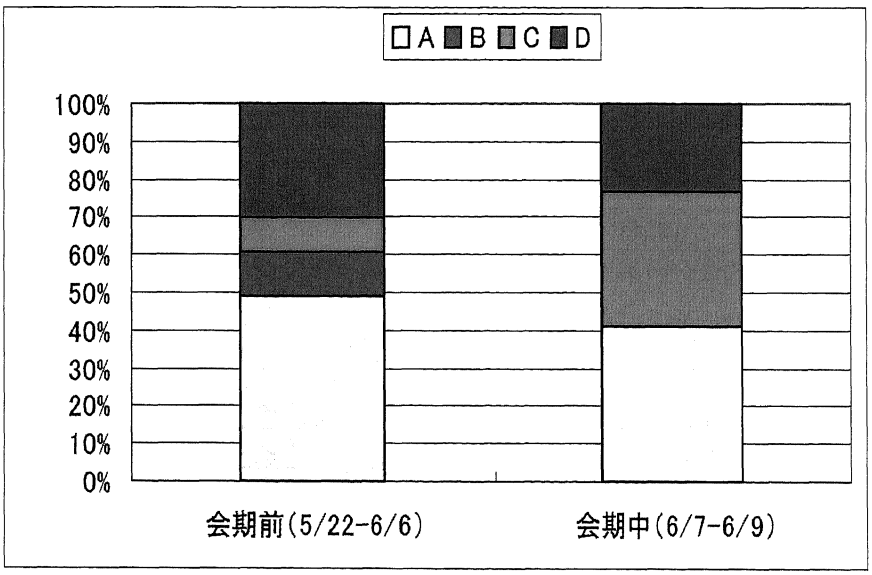

図 4.12: ユーザが選択した足跡リンクを推薦するために用いられた推薦アルゴリズムの割合

これによると，会期前は推薦アルゴリズムAによって推薦されたリンクが足跡リンクとして選ばれている．これは，会期前において，推薦アルゴリズムAによる推薦リンクが選

ばれていたことと関係あるものと思われる（図 4.5）．一方で，会期中になると，会期前とは違って推薦アルゴリズム C によって推薦されたリンクが足跡リンクとして選ばれている．したがって，足跡リンクにおいても，ユーザがマーキングを付与した語を利用した推薦リンクが選ばれたことが分かった．

4.3.3 tfidf 値の高い語が情報探索に有益かどうか

3 章において，ユーザが下線を付与した語は tfidf 値の高い語が多いことが分かった．そこで，tfidf 値が高い語が情報探索に有益なのかどうかについて調べた．

まず初めに，ユーザがマーキングを付与した語の tfidf 値について調べた．表 4.4 は発表ページに含まれる語と，会期前においてマーキングが付与された文字列に含まれる語の tfidf 値群の中央値と四分位偏差値，語数を示している．表 4.5 は発表ページに含まれる語と，会期中においてマーキングが付与された文字列に含まれる語の tfidf 値群の中央値と四分位偏差値，語数を示している．

表 4.4: 発表ページと会期前においてマーキングされた文字列に含まれる語に関する tfidf 値

	発表ページ	マーキング文字列
中央値	4.3	5.4
四分位偏差値	2.6	3.2
語数	6404	76

表 4.5: 発表ページと会期中においてマーキングされた文字列に含まれる語に関する tfidf 値

	発表ページ	マーキング文字列
中央値	4.3	5.0
四分位偏差値	2.6	4.5
語数	6404	52

発表ページには，発表タイトルと発表概要文および第一著者名が書かれている．分析に使用した語は茶筌による形態素解析の結果，名詞と未知語と判断された 2 文字以上の語である．ただし，人名の場合のみ，一文字も使用した．発表ページに含まれる語群の tfidf 値とマーキングが付与された語群の tfidf 値の分散に差がないかどうかについて調べた．アン

サリ・ブラドレイ検定を行った結果，$p > 0.05$ となり，両群の分散に有意差がないということが分かった．次に発表ページに含まれる語群の tfidf 値とマーキングが付与された語群の tfidf 値とで中央値の差がないかを調べるために，マンホイットニーの U 検定を行った．その結果，会期前は U=172722.5($p < 0.05$)，会期中は U=120519($p < 0.05$) となり，両方の期間において，発表ページに含まれる語群の tfidf 値とマーキングが付与された語群の tfidf 値の中央値の差が有意であるという結果が得られた．

図 4.13 は発表ページに含まれる語の tfidf 値のヒストグラムで，図 4.14 は会期前におけるマーキングに含まれる語の tfidf 値のヒストグラムである．図 4.15 は会期中におけるマーキングに含まれる語の tfidf 値のヒストグラムである．

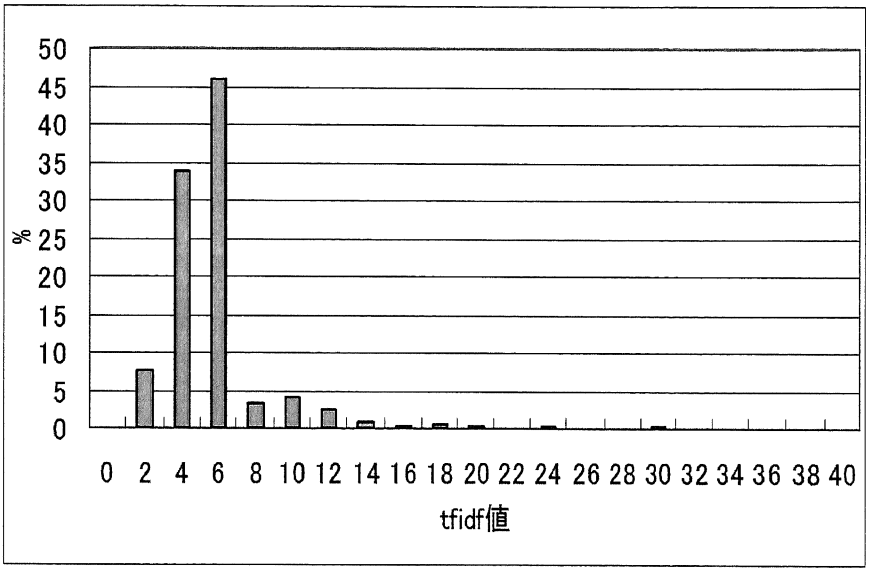

図 4.13: 発表ページに含まれる語の tfidf 値のヒストグラム

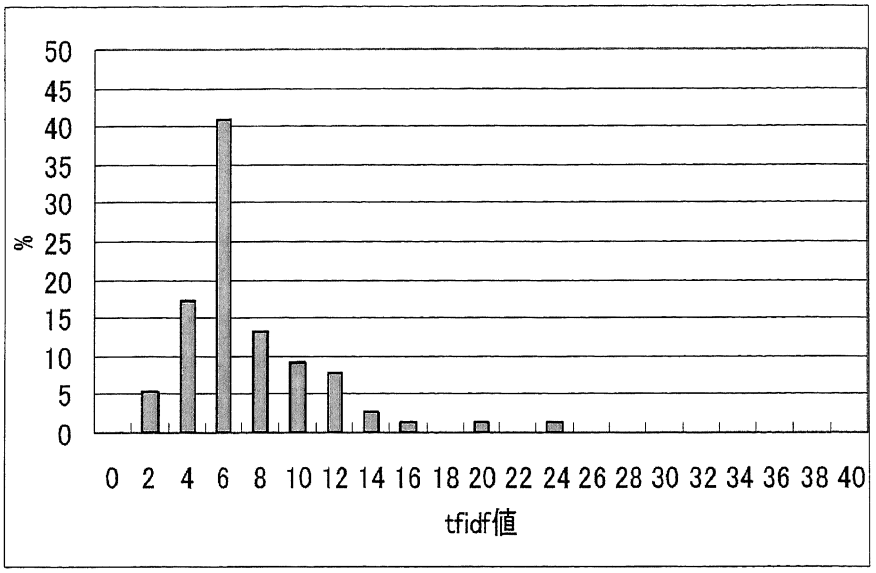

図 4.14: 会期前におけるマーキング文字列内の語の tfidf 値のヒストグラム

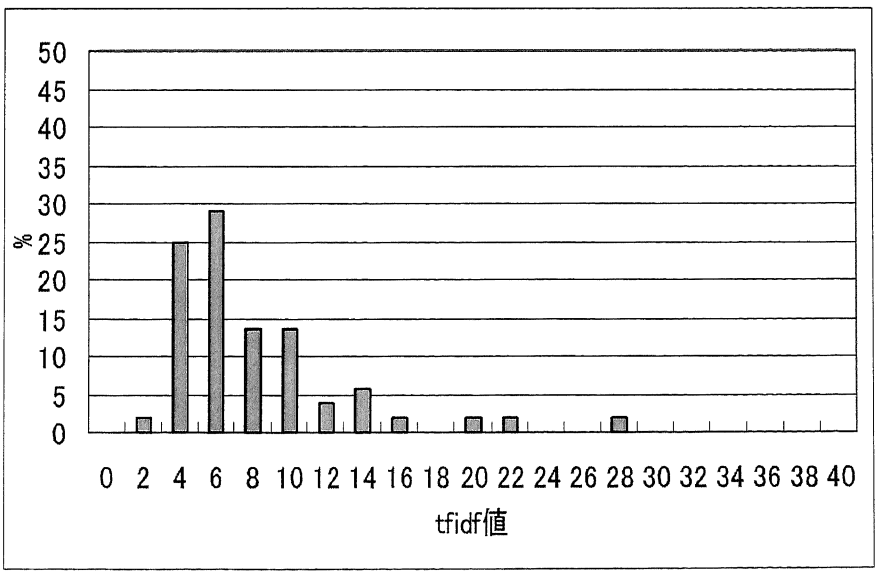

図 4.15: 会期中におけるマーキング文字列内の語の tfidf 値のヒストグラム

発表ページに含まれる語の tfidf 値 6〜7 の区間における語の出現頻度が最も高く，8 以

上の区間は語の出現頻度が低い．一方で，会期前および会期中におけるマーキングが付与された語の tfidf 値の 8 以上の区間は，発表ページの語の tfidf 値の同じ区間よりも語の出現頻度が高い．したがって，マーキングが付与された語には tfidf 値の高い語が多く含まれることが分かった．また，会期前においてマーキングが付与された語群の tfidf 値と会期中においてマーキングが付与された語群の tfidf 値とで中央値に差があるかどうかを調べると，マンホイットニーの U 検定の結果，U=1926.5($n.s.$) となり，両者に有意差はなかった．これらの結果より，3 章と同様に，ユーザがマーキングを付与した語には tfidf 値の高い語が多く含まれることが分かった．

次に，マーキングが付与された語の tfidf 値の高い語が情報探索に役立ったのかについて調べた．合口では，ユーザが推薦リンクをクリックすると，ユーザが選択した文字列にマーキングを付与した．このマーキングを付与した文字列のことを，マーキング文字列と読んでいた．図 4.16，4.17 は，会期前と会期中において，マーキング文字列内の語の tfidf 値とユーザが選択した推薦リンクを算出するために用いられた推薦アルゴリズムによるヒストグラムである．

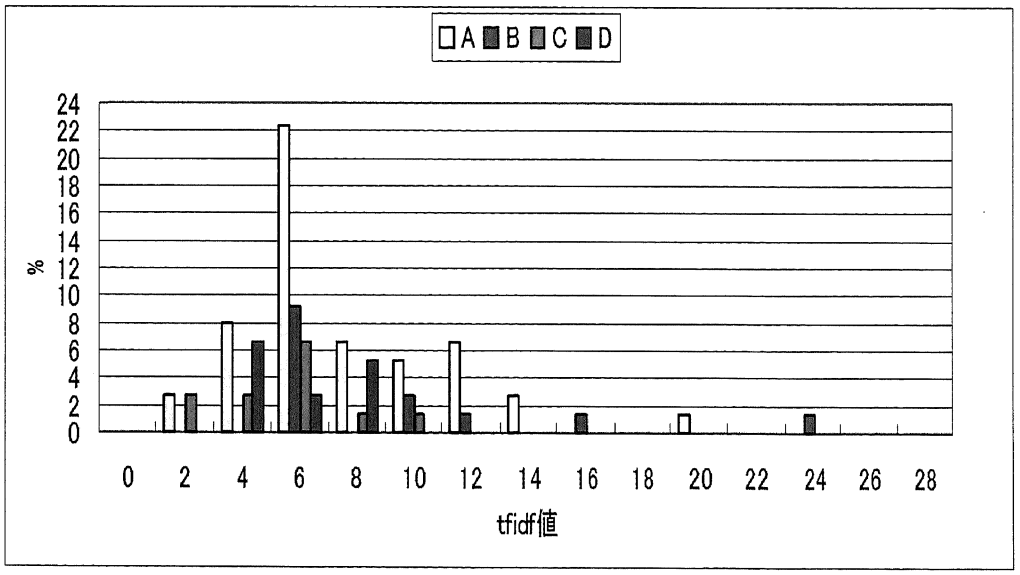

図 4.16: マーキング文字列内の語の tfidf 値とユーザが選択した推薦リンクを算出するために用いられた推薦アルゴリズムによるヒストグラム（会期前）

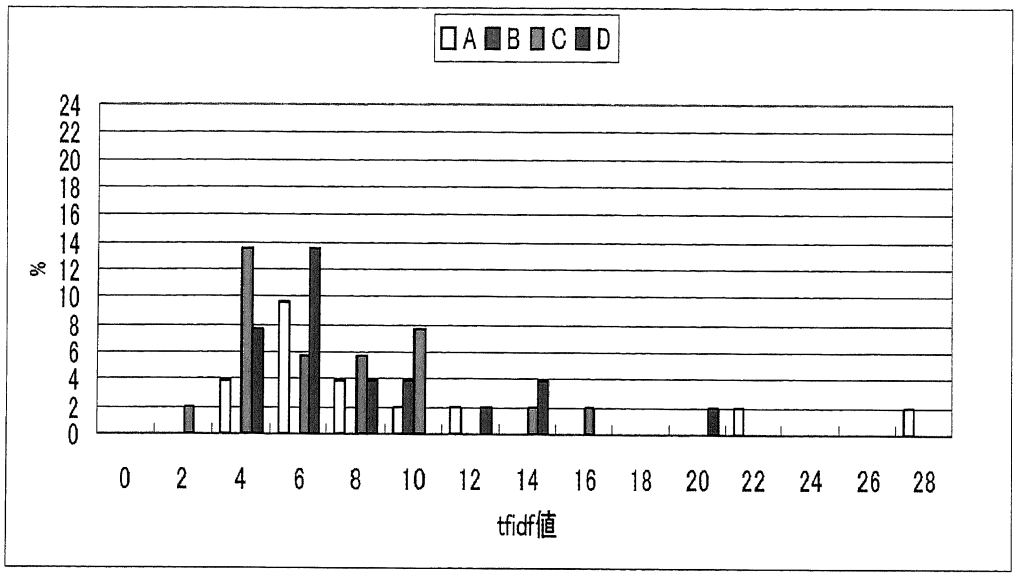

図 4.17: マーキング文字列内の語の tfidf 値とユーザが選択した推薦リンクを算出するために用いられた推薦アルゴリズムによるヒストグラム（会期中）

図 4.16 より，会期前は tfidf 値が 2〜15 の区間において，推薦アルゴリズム A による推薦リンクを選択することによって付与されたマーキング文字列内の語の出現頻度が高い．推薦アルゴリズム A は発表ページ内に含まれる全ての語を利用するため，tfidf 値の低い語から高い語まで，どれを選択しても何らかの推薦リンクを算出することができる．また，図 4.5 で示したように，ユーザは推薦アルゴリズム A による推薦リンクを最も選択していたため，幅広い区間において語の出現頻度が高くなったものと思われる．会期中では，tfidf 値が 4〜5 の区間において，推薦アルゴリズム C による推薦リンクを選択することによって付与されたマーキング文字列内の語の出現頻度が高い．会期中におけるマーキング文字列に含まれる語の tfidf 値の平均値は 7.0 なので，平均よりも tfidf 値が低い語の出現頻度が高いといえる．推薦アルゴリズム C は発表ページ上でユーザが選択した文字列と他のページに付与されているマーキング文字列とのマッチングによって算出される．したがって，ユーザが選択した文字列内の語が他の発表ページに含まれていないと推薦されることはない．語が多くの文書に含まれる場合，tfidf 値は低くなるため，推薦アルゴリズム C では tfidf 値の低い語を利用することが多くなるために，tfidf 値の低い区間に推薦アルゴリズム C による推薦リンクによって付与されたマーキング文字列内の語の出現頻度が高くなっ

たのだろう．実際に，各推薦アルゴリズムによる推薦リンクの選択よって付与されたマーキング文字列内の語の tfidf 値の平均値を求めたところ，表 4.6 のようになり，会期前および会期中の両期間において，推薦アルゴリズム C による推薦リンクの選択によって付与されたマーキング文字列内の語の tfidf 値の平均値が最も低いことが分かる．

表 4.6: 推薦アルゴリズムによる推薦リンクの選択によって付与されたマーキング文字列内の語 tfidf 値の平均値

	A	B	C	D
会期前	6.6	8.6	4.5	4.8
会期中	8.6	8.4	6.0	6.7

一方で，会期前および会期中において，推薦アルゴリズム B による推薦リンクの選択によって付与されたマーキング文字列内の語の tfidf 値の平均値が高い．推薦アルゴリズム B はユーザのページへの評価値をページに付与されたマーキング文字列の数としている．したがって，ユーザが選択した文字列内の語が他のページに含まれていなくても推薦されることがある．tfidf 値の高い語は，他のページに出現する頻度が少なく，かつ該当ページに多く含まれる語のことであるから，推薦アルゴリズム C や D によって tfidf 値の高い語が含まれるページを推薦されることは少ない．推薦アルゴリズム B では，tfidf 値の高い語が他のページに含まれていなくても，ページを推薦される機会が多いため，マーキング文字列内の語の tfidf 値の平均値が高くなっている．ちなみに，推薦アルゴリズム B による推薦リンクの選択によって付与されたマーキング文字列内の語のうち，最も tfidf 値の高い語である「モーフィング」は一つの発表ページにしか出現しない．4.3.1 より，ユーザは会期中において，推薦アルゴリズム C による推薦リンクを最も選んでいることから，ユーザは tfidf 値の高い語とは関係なく，他人が付与したマーキング文字列内の語を利用した情報探索を好んだといえる．

4.4 考察

本章では，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際，情報探索に役立つのかについて調べた．

4.3.1 より，ユーザはページ間類似度の高い発表ページよりもマーキングが付与された文字列内の語を利用した推薦アルゴリズム C に基づいて推薦された発表ページを好むこと

が示唆された．とくに学会期間中においては，ユーザが文書内で注目した語に関するページを探すという目的を持って使用したため，マーキングが付与された文字列内の語を利用したアルゴリズムによる推薦ページを好む傾向があったものと思われる．推薦アルゴリズム C では，ユーザが発表ページ内で選択した語が他の発表ページ上でマーキングされていないと，推薦されない．システムの利用者は，人工知能学会全国大会に参加した人であり，対象としたページは学会で発表される論文の概要文等が記述された発表ページである．したがって，今回のような結果が得られたのは，ドメインに特化したコンテンツ集合に対して，共通の興味をもった複数ユーザがマーキングを付与したから，と考えられる．このような状況下において，対象となるコンテンツやユーザの数が増えれば，マーキングを利用した情報探索がさらに有用になるものと思われる．表 4.7 は，推薦アルゴリズム C によって推薦されたページがユーザによって選択されたときに，ユーザが発表ページ上で選択した文字列と他の発表ページに付与されているマーキング文字列間でマッチングした語である．この語集合は，合口を利用したユーザ間で共有された語集合といえる．

表 4.7: 推薦アルゴリズム C における選択文字列とマーキング文字列間のマッチング語

リング，データ，利用，Web，武田，複数，共有，知識，ネットワーク，クラスタリング，システム，ウェブ，情報，関心，ActionLog

4.3.2 における分析では，ユーザがマーキング文字列から他の発表ページへ張ったリンクが好まれていたことが分かった．これに関しても，共通の興味を持ったユーザ，ドメインに特化したコンテンツという実験環境が影響していたものと思われる．

4.3.3 の分析では，全ユーザがマーキングを付与した文字列には tfidf 値の高い語が多く含まれることが分かった．しかし，マーキング文字列内の語の tfidf 値が低い語を利用した推薦アルゴリズム C による推薦が好まれていた．tfidf はコンテンツ内の特徴度を計算するための指標であるため，tfidf 値の高い語はユーザがコンテンツを見たときに，コンテンツの内容を把握するときに有用である．表 4.7 のマッチング語のうち，「システム」の tfidf 値は 1.2 で，「データ」の tfidf 値は 2，「共有」の tfidf 値は 2.6 である．これらは，表 4.6 で示した tfidf 値の平均値よりも低い．したがって，tfidf 値の低い語を利用した推薦アルゴリズム C による推薦ページが好まれていたということは，tfidf 値の低い語でも，複数ページ間を連結する役割を担うメタデータとして利用できるといえる．

4.5 まとめ

本研究では，ページ間類似度やマーキングされた文字列内の語や文字列を使ったページ推薦を行う合口を運用して，マーキングが付与された語が情報探索に有益かどうかついて調べた．その結果，ユーザは学会期間中において，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．また，ユーザはシステムによって推薦されたページよりも，他人がマーキング文字列から他のページへ張ったリンクを選択することも示唆された．さらに，ユーザが付与したマーキング文字列内の語のうち，tfidf 値の高い語が情報探索に有益かどうかについて調べたところ，ユーザは tfidf 値の高い語とは関係なく，他人が付与したマーキング文字列内の語を利用した情報探索を好むということが分かった．これらの結果より，他人がマーキングによってページ内で注目した語は情報探索に有益であることが見出された．

第5章

メモとコンテンツ内容との関連性に関する分析

本章では，ユーザがWebコンテンツに付与したメモからコンテンツの内容と関連のある語が獲得できるのかについて調査した．本研究では，学会発表時に聴講者が2種類のメモを書くことができるシステム“memoQ”の第21回人工知能全国大会（JSAI2007）における運用で得られたデータを使って，ユーザが論文発表に対して入力したメモ内の語が発表内容と関連があるのかについて調査した．

5.1 はじめに

本研究では，Webコンテンツ対して，コンテンツの内容と関連のある語が記述された意味的メタデータを生成することを目標としている．そこで，コンテンツに含まれない語のうち，コンテンツの内容と関連のある語を獲得するために，ユーザのメモ書きに着目した．しかし，ソーシャルブックマークのタグのように，ユーザが自由にコンテンツにキーワードを付与すると，ユーザの主観が反映されるという問題があった．そこで本研究では，学会の発表論文に対して意味的なメタデータを生成するために，ユーザが発表聴講時にメモをとるときの状況に着目した．ユーザが発表を聴講する際に書くメモは，発表内容に関して記述するメモと，発表内容に関する感想などのユーザの主観が反映されたメモの2種類がある考えられる．そこで本章では，発表時に2種類のメモを入力することができるシステム“memoQ”の運用で得られたデータを使って，それぞれのメモから発表内容と直接関係のある語とユーザの主観が反映された語を獲得することができるかどうかについて調査した．

5.1.1 関連研究

学会において，発表聴講時にメモを入力できるシステムがいくつか開発されている．Lock-on-Chat[西田 06] では，学会発表中に聴講者間でコメントを共有できるシステムである．発表スライドを撮影して Web サーバにアップロードし，複数の発表スライド画像内の好きな箇所にコメントを付与できる．また，コメントは別途，時系列に表示されるため，チャットのようなシステムになっている．ActionLog[沼 07] では，発表を聴講するときに Weblog 形式で記事を投稿することができる．また，投稿した記事は他のユーザ間で共有されるため，発表に関して投稿された記事の一覧を見ることができる．これらのシステムは，発表に対してユーザが自由にコメントを入力できるようになっているため，コメントにはユーザの主観が反映された語が含まれる可能性がある．本研究では，発表聴講時のコンテキストを利用することによって，ユーザの主観が反映された語が含まれるメモと発表内容に関する語が含まれるメモの 2 種類を獲得することを目指す．

5.2 memoQ

memoQ は発表の聴講を支援する目的で開発したシステムである（図 5.1）．学会発表において，発表後に質問が一つも出ないために議論されずに終わってしまうことがある．また，特定の聴講者ばかりが質問し続けることによって，多角的な議論がなされない，という問題もある．発言をしない聴講者の中には，本当は質問をしたいと思っているが，恥ずかしがって質問するのをためらってしまう人や，公表するほどの質問でもないと判断してしまう人がいることが考えられる．しかし, 発言されなかった質問の中には，発表者および聴講者の双方にとって有益な質問が埋もれてしまっている可能性がある．そこで，学会発表において聴講者が質問を出しやすくし，発表者および聴講者間で質問を共有することで質疑応答の場を活性化することを目的としたメモ書きシステム “memoQ” を提案した．学会において，発表の聴講者はノートにメモを取りながら聴講することがある．聴講者が書いたメモには発表に対する意見や疑問が含まれるため，他者と共有することにより，議論の材料として活用できるものと思われる．

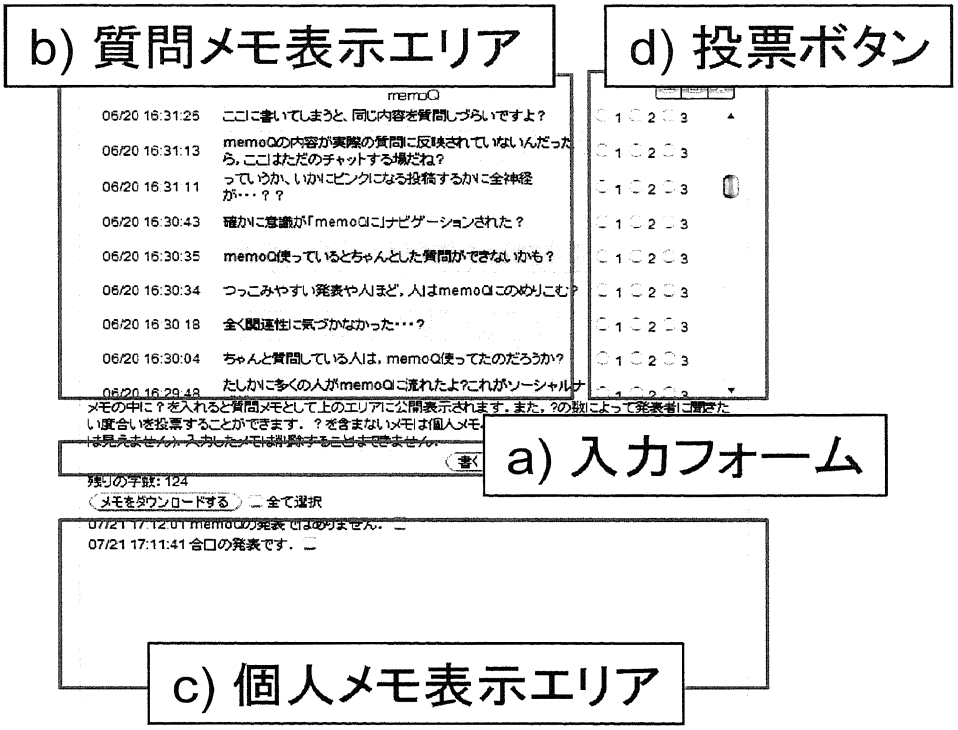

図 5.1: memoQ のインタフェース

5.2.1 デザイン設計

memoQ のデザイン目標は，学会発表において聴講者が質問を出しやすくし，発表者および聴講者間で質問を共有することでコミュニケーションを促進させることである．この目標を達成するため，以下の点に注目した．1 つ目は，思ったことを素早く書き込めるようにすることである．熟考した上での質問より，発表を聴講しているときに，ふと疑問に思ったことの中に有益な質問が含まれているかもしれない．2 つ目は，質問のメモを書きやすくすることである．他人に公開されることがいやな人でも書きやすくする必要がある．3 つ目は，発表に集中させることである．システムは発表を聴講するときの補助の役割を果たすものなので，聴講者がシステムを使うことに熱中することは望ましくない．4 つ目は，聴講者が書いた質問の中から重要度の高い質問を決定することである．聴講者がたくさん質問を記入した場合，全ての質問に発表者が答えるのは困難である．

本システムでは，聴講者が思い付いたことを素早くメモ書きできるようにするために，現

在WWW上で多くの人に利用されているTwitter[1]のインタフェースを採用した．Twitterでは，ユーザが“What are you doing?”という問いかけに答える形で，140文字制限でメッセージを入力することができる．発表を聴講しているときに，時間の流れに沿って自然と思いついた意見や疑問をメモとして残す場合，長文を書くことは想像しにくい．そこで本システムにおいても，短いメモだけを入力できるようにするためにTwitterの入力インタフェースを採用する（図5.1(a)）．

聴講者が入力したメモは，他人が発表に対してどのような質問があるのかを知るために，ユーザ間で共有されることが望ましい．また，共有することで，同じ内容の質問がかかれるのを防ぐことができる．しかし，全てのメモを共有した場合，質問のためのメモだけでなく，聴講者の発表に対する感想や意見といったメモが含まれる可能性があるため，質問を捜すのが難しくなるかもしれない．また，聴講者間でメモをリアルタイムに共有すると，聴講者間で会話が生じてしまい，質問が出にくくなるということが想定できる．そこで本システムでは，入力されたメモの中に“?”が含まれているときは，質問メモとして匿名で他の聴講者に公開し(図5.1(c))，その他のメッセージは個人メモとして非公開にする．これにより，聴講者は問いかけに対する返答ができなくなるため，質問を出すことに集中できる．また，匿名でメモを表示することによって，質問を公表することに恥ずかしさを感じる人でも入力しやすいようにしている．聴講者が入力したメモは，全て個人メモとして一覧表示されるようになっている(図5.1(b))．聴講者が入力したメモは，全て時系列で一覧表示する．システムの利用者には，あらかじめメモの中に“？”が含まれていたら質問メモとして公開することを知らせる．

本システムは，質問メモに対して聴講者が発表者に聞きたいと思う度合いを投票する機能を提供する．聴講者が入力した質問メモの中には，他の聴講者も聞きたいと思う質問もあれば，それほど聞きたくない質問も含まれる．そこで，投票機能を提供することにより，聴講者間で質問メモの重要度を決定できるようにする．本システムでは，発表者に聞きたい度合いを質問メモの横にある投票ボタンをクリックすることで投票してもらう．評価は一つの質問メモにつき，三段階評価で投票してもらう．システムは聴講者が記述したメモの中の？の数を聴講者の投票として扱う．従って？が1つのときは1という評価を，3つのときは3という評価をしたとする．聴講者は投票ボタンによって後から投票数を変えることができる．また，聴講者が自分あるいは他人の質問メモの投票ボタンをクリックすると，質問メモの背景色を濃くする（図5.1(d)）．背景の色は，点数に応じて黄色，薄いピ

[1] http://twitter.com/

ンク，濃いピンクと変化させる．色の変化は聴講者間で共有されるため，リアルタイムにどの質問に人気があるのかを知ることができる．自分が書いた質問メモが投票によって支持されることによって，ゲーム感覚で質問を書きやすくなるのではないかと思われる．

本システムは，発表直後の質疑応答の時間に活用するため，聴講者が投票した結果を利用して質問メモのまとめ図を表示する（図 5.2）．まとめ図は，投票数の多い順に質問メモを上から並べて表示する．この図を発表者や座長が見ることで，聴講者からどういった質問があるのかや，重要度の高い質問を把握することができる．

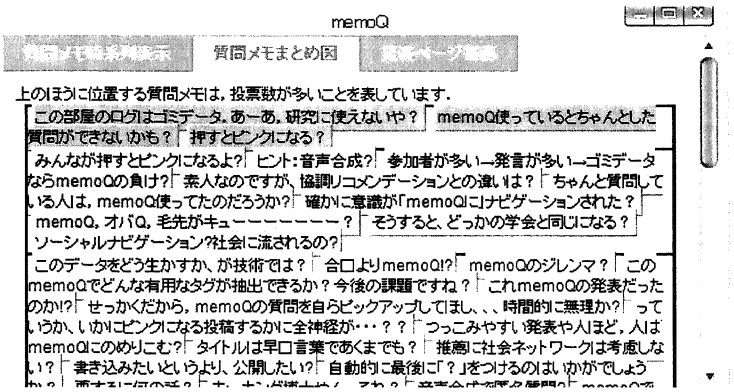

図 5.2: 質問メモのまとめ図

5.2.2 操作方法

memoQ では，発表ごとに用意されている発表ページ上にメモを入力することができる．ユーザは大会支援システムにログインした後，発表ページ上にある memoQ のスタートボタン（図 5.3）をクリックすることによって，memoQ を起動することができる．ユーザが「memoQ：発表者に質問する！」と書かれたボタンをクリックすると，図 5.1 に示した memoQ がポップアップ表示される．

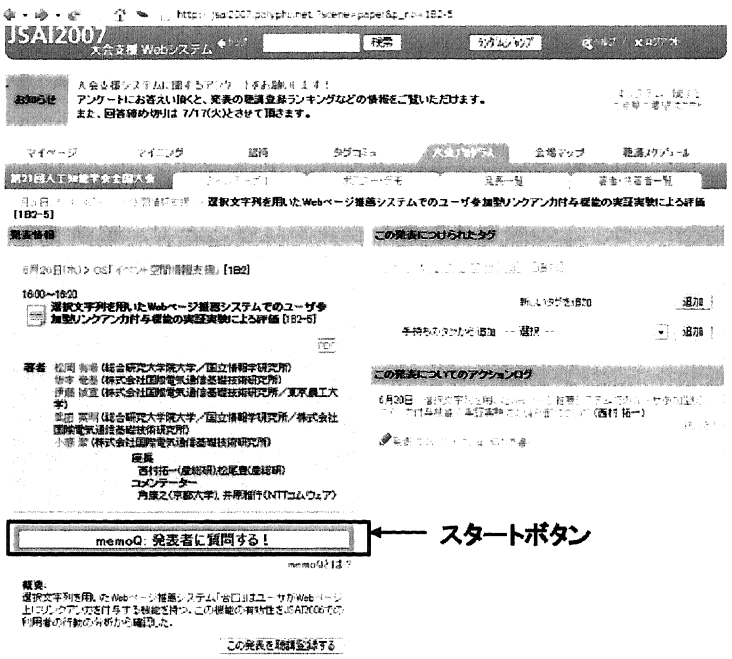

図 5.3: 発表ページと memoQ のスタートボタン

5.2.3 運用結果

memoQ を 2007 年 6 月 20〜22 日に開催された JSAI2007 で運用した．JSAI2007 では，全部で 61 のセッションがあり，330 件の発表があった．発表中に本システムを利用したユーザは 72 名いた．そのうち，質問メモあるいは個人メモを書いたユーザは 61 名で，投票だけしたユーザは 11 名だった．また，ユーザが入力したメモの総数は 2556 行あり，質問メモは 1850 行，個人メモは 706 行だった．メモが付与されたセッション数は 45 で，発表数は 148 であった．データは発表時間内に付与されたものに限る．

5.2.4 入力インタフェースの効果

memoQ では，ユーザが思ったことを素早く入力できるように最大 140 文字（日本語は一文字が 2 バイトのため最大 280 バイト）まで入力できるインタフェースを用意した．この効果を確かめるために分析を行った．聴講者によって入力されたメモの平均文字バイト数は，40.0 バイトだった．図 5.4 のメモのバイト数によるヒストグラムより，20〜30 バイ

トのメモの数が多いことが分かる．このように，ユーザは短いメモを入力する傾向があったため，発表聴講時に思いついたことをすばやく入力していたものと思われる．

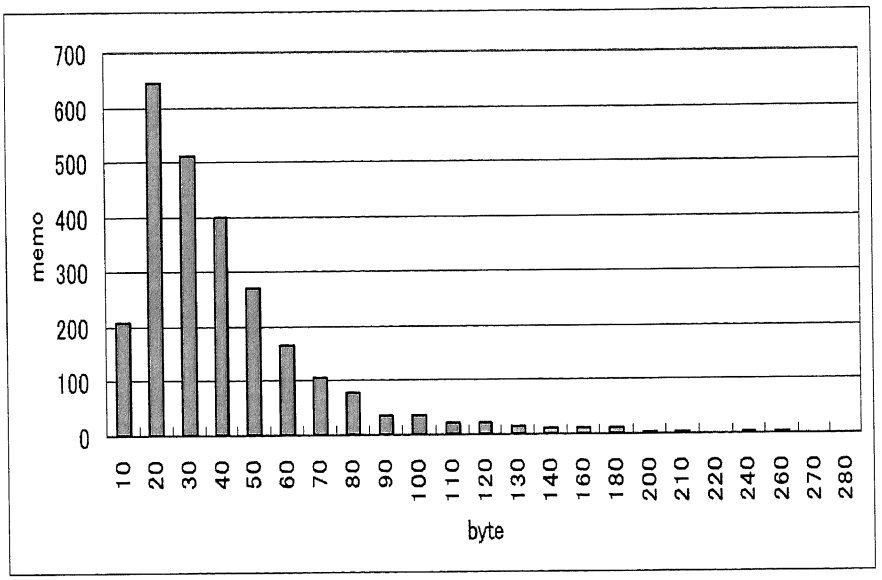

図 5.4: メモのバイト数によるヒストグラム

5.2.5 質問の出しやすさに関する分析

memoQ は，メモを公表することに対して恥ずかしいと思う人のために匿名で “?” が含まれるメモのみを質問メモとして公開した．そこで，質問メモの数について調べたところ，ユーザが入力したメモのうち，質問メモの割合は 72.3%だった．また，各ユーザの質問メモと個人メモの使われ方に関して調べると，質問メモの数が個人メモの数よりも多いユーザの割合は 72.1%（44 名/61 名）だった．従って，本システムの利用者は質問メモを書くために使ったユーザが多いということが分かる．各発表で入力された質問メモの数と個人メモの数，システムの利用者数を調べた．図 5.5 によると，質問メモの数が多い発表では，個人メモの入力数が少ない．一方で，図 5.6 より，質問メモの数が少ない発表では，個人メモの入力数が少ない．利用者数に注目すると，質問メモの入力数が多いほうが少ないよりも利用者数が多い．従って，ユーザの参加者が多いほど質問メモが書きやすい傾向があることが分かった．

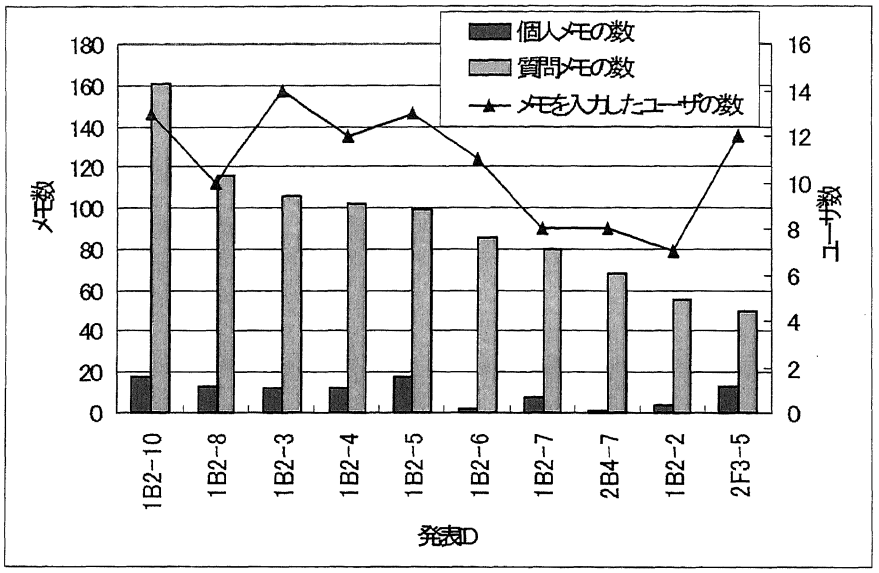

図 5.5: 質問メモの数が多い順に上位 10 個の発表で入力された質問メモと個人メモの数と利用者数

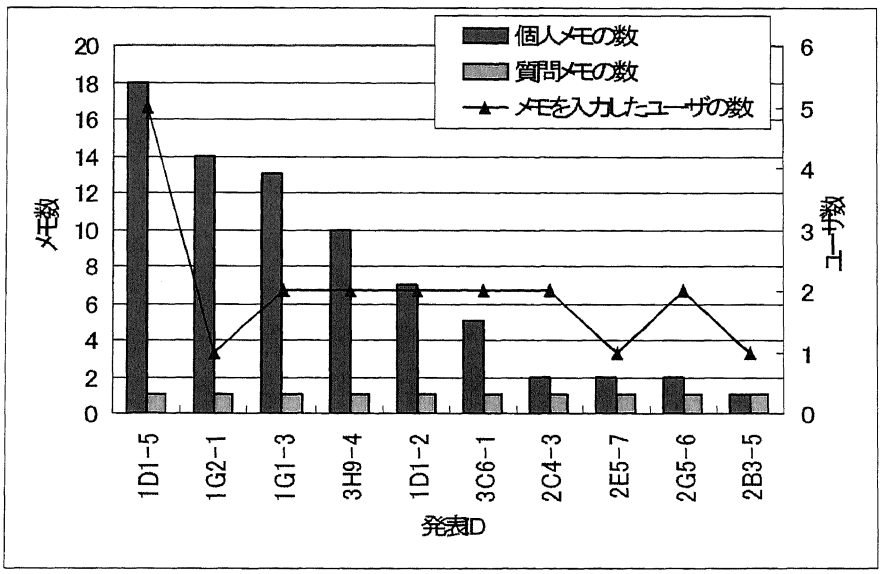

図 5.6: 質問メモの数が 1 の発表で入力された質問メモと個人メモの数と利用者数

5.2.6 発表に集中したかどうかの分析

memoQ では，発表に集中してもらうために，“?” が含まれる質問メモだけを他者に公開することで会話を発生させないような工夫をした．そこで，ユーザはどのような目的で質問メモを入力したのかについて調べた．質問メモに書かれている文章のうち，目視で発表内容と関係のあるメモかそうでないメモかについて調べた．質問メモに書かれている内容は，下記の分類ができる．

1. 発表に対する質問・感想・意見・ノート
2. 実際にあった質疑応答のメモ
3. 発表内容と関係のないメモ

1 の分類は，発表に関連する内容が書かれたメモが当てはまる．語尾に「〜か?」や「〜では？」,「〜と思います」などと書かれていれば，質問や感想，意見であると判断した．また，メモの先頭に?が書かれていたり,「メモ?：」と明示した上で書かれているメモは，発表のノートをとっているものと判断した．2 の分類は，実際に会場で発生した質疑応答が書かれたメモが当てはまる．メモに,「【質問コピー】」や「[フロアからの質問]」と書かれているものを指す．3 の分類は，発表内容と関係のないメモのことであり，雑談が当てはまる．これらの分類方法によって質問メモを分類した結果，1 に分類されたメモは 1112 個，2 に分類されたメモは 27 個，3 に分類されたメモは 711 個あった．当初は，発表内容に関することが質問メモとして入力されることを期待していたが，発表内容と関係のないメモが 38.4%あった．図 5.7 は，セッションごとに，発表内容と関係のない質問メモと関係のある質問メモの数を示している．

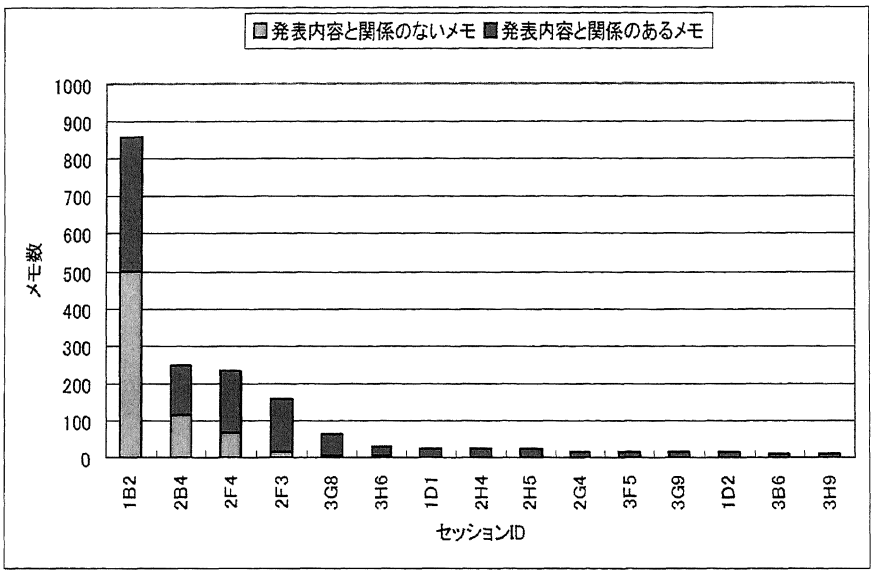

図 5.7: セッションごとの発表内容と関係のある質問メモと関係のない質問メモの数

ここで，発表内容と関係のある質問メモは 1 と 2 に分類されたメモのことを，発表内容と関係のないメモは 3 に分類されたメモのことを指している．入力された質問メモの数の多い順に，上位 15 個のセッションによる図である．これによると，質問メモが多く入力されたセッションでは，発表内容と関係のないメモが多く入力されたことが分かる．質問メモを入力したユーザを調べたところ，発表内容と関係のないメモの数が多いセッション 1B2 と 2B4 で質問メモを入力したユーザ（1B2 は 22 人，2B4 は 16 人）のうち，両方で質問メモを入力したユーザは 11 人いた．セッション 1B2 および 2B4 において，入力された発表内容内容と関係ないメモのうち，これら 11 人が入力した割合は，1B2 で 72.8%，2B4 で 86.2%だった．このように，特定ユーザが発表内容と関係のないメモを入力する頻度が高いことがわかった．図 5.7 より，発表内容に関するのメモで占められるセッションもあるため，特定のユーザを除いては，想定していた使い方をしているユーザがいることがわかる．従って，特定のユーザを除いては発表に集中していたといえる．質問メモは匿名で公開していたため，特定ユーザが質問とは関係のない雑談を自由にしていたものと思われる．

5.2.7 投票に関する分析

memoQ では，質問メモの重要度を決めるために，自分あるいは他人の質問メモに対して聞きたい度合いを投票することができた．他人の質問メモに 1 回でも投票したことのあるユーザは 58 名おり，平均回数は 15.3 回だった．一人につき，一番多い投票回数は 151 回もあった．次に，投票数が多い質問は発表内容と関係のあるメモかそうでないメモなのかを調べたところ（図 5.8），投票数が高いと，発表内容と関係のないメモの割合のほうが高いことが分かった．これは，投票数が多かったのは，発表内容と関係のない質問メモだったため，ユーザが真面目に投票しなかったものと思われる．従って，ユーザは投票機能を積極的に使っているが，ユーザは発表内容と関係のある質問に投票するよりは，質問の内容が面白いかどうかによって投票していたことが分かった．

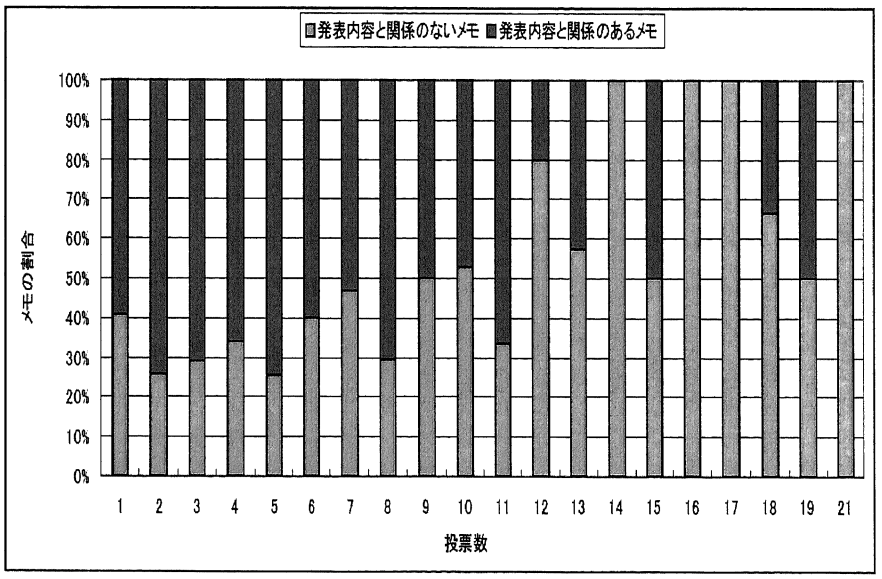

図 5.8: 投票数ごとの発表内容と関係のある質問メモと関係のない質問メモの割合

5.2.8 アンケート結果

学会の終了後，memoQ に関するアンケートを行ったところ，27 名の回答者がいた．アンケートの文面は下記のとおりである．

Q1. memoQ は発表の聴講時に役立ちましたか？

Q2. 質問メモは匿名で公開されるので，入力しやすかったですか？

Q3.　他の方が入力した質問メモをリアルタイムに見ることは，ご自身にとってどういう影響がありましたか？

Q4.　質問メモのまとめ図は役に立ちましたか？

Q1 は memoQ の評価を調べるための質問である．回答は，1 から 5 点でそれぞれ「まったく役に立たなかった」「あまり役に立たなかった」「どちらともいえない」「役に立った」「大変役に立った」の 5 段階からの選択である．アンケートの結果によると，1 点が 2 人，2 点が 3 人，3 点が 4 人，4 点が 13 人，5 点が 5 人で，平均得点は 3.6 点だった．低い評価をしたユーザは，「システムを起動できなかった」，「存在が分かりづらかった」，「使い方が分からなかった」と意見していたので，システムを使用する前の問題により，システムの評価が低くなったことが分かった．

Q2 は匿名の効果について調べる為の質問である．回答は，1 から 3 点でそれぞれ「入力しにくかった」「どちらともいえない」「入力しやすかった」の 3 段階からの選択である．結果は，1 点が 1 人，2 点が 14 人，3 点が 12 人で，平均点は 2.4 点だった．このように，「どちらともいえない」というユーザのほうが「入力しやすかった」というユーザよりも多かった．しかし，「入力しにくかった」というユーザは 1 人だけだったので，匿名にすることでメモが書きにくいと感じるユーザは少なかったことが分かる．

Q3 は質問メモの共有の効果について調べる為の質問である．ここでは，ユーザに自由に回答を記述してもらった．好意的な意見としては，「他の人がどのように感じながら発表を聞いているのか知ることができ，疑問の共有や解決に役に立った.」，「質問メモをみることによって新しい質問を思いつくという創発効果があった.」，「自分にはなかった視点に気づけた.」，「同じ質問を持っている人などがいることが分かり良かった.」というものがあった．これらの意見から，質問を共有する効果があったことが分かった．一方で，「面白いがプレゼンターへの意識は薄れる」，「ちゃんとした質問は参考になったが、ただのチャットになっているときには，聴講に集中できなくなった.」という意見もあり，発表に集中できないときがあったことが指摘されている．これは，5.2.6 において発表と関係のないメモが書かれたセッションがあったことと関係があるものと思われる．

Q4 は投票機能の効果について調べる為の質問である．回答は Q1 と同じ 5 段階評価に加えて，「まとめ図の機能があることを知らなかった」という回答を用意した．27 名のうち 2 人が無回答で，6 人が「まとめ図の機能があることを知らなかった」を選択していた．その他のユーザによる回答は，1 点が 2 人，2 点が 7 人，3 点が 3 人，4 点が 6 人，5 点が 1 人で，平均点は 2.8 点だった．このように，あまり良い評価でなかったのは，5.2.7 で示した

ように，発表内容と関係のない質問の点数が高くなったため，真面目な質問が上位に来なかったからだと思われる．

最後に自由に感想や意見を書いてもらった．その中に「みんなが投票すると色が濃くなるのが面白く，みんなが投票してくれるような質問を考えていた」という意見があったので，投票すると質問メモの背景色が変化させる効果があったことが分かった．一方で「匿名にすると，無責任な発言が多くなるので，できれば匿名でない方がいいと思いました.」，「難しいとは思いますが、誹謗中傷につながるコメントがなるべく少なくなるようにするとよいと思います」,「質問の状況を発表者に見えるようにするか，利用者同士の議論もできるようにしておいてもらえないとつまらない．発表の向上のためにも，もっと自由なシステムの方が良かったのでは」という問題点を指摘する意見があった．

5.2.9 考察

本システムでは，学会において聴講者から質問を出しやすくするために，質問のメモのみを匿名で公開した．匿名性の効果については [Joinson 99] によって，コミュニケーションの敷居の高さを下げる効果があることが分かっている．本システムの運用によって，質問メモのほうが個人メモよりも多かったことから，その効果が証明された．しかし，中には発表内容と関係のないメモが多く入力されたセッションがあった．これは特定のユーザが，発表内容と関係のないメモを多数入力していたからである．質問メモの入力傾向をみると，一人が不真面目なメモを書き込むと同調するという傾向が見られた．これは時系列でメモを表示させていたため起こったことと考えられる．本研究では意味的メタデータを生成するときに集合知を活用することを提案した．集合知では多様性，独立性，分散性，集約性の四つの要件がそろったときに機能するとされている．しかし，質問メモを時系列で表示することによって，ユーザの意見が悪い方向に流されるという現象が生じてしまった．したがって，memoQ は集合知を上手く活用できているとはいえない．今後の設計方針としては，雑談が発生するのを防ぐために，メモの表示方法を時系列ではなく，質問メモの内容に合わせて分類するといった工夫が必要がある．その他の解決方法としては，アンケートに記述されていたように，質問の状況を前のスライドで映すことによって発表者も見えるようにすれば，発表内容と関係のないメモが少なくなるかもしれない．また，投票機能においても，発表内容と関係のないメモに点数が集中してしまった．これに関しては，マイナス評価ボタンを用意することで投票数を減少できる可能性がある．

5.3 分析

本節では，発表聴講時のコンテキストに応じて入力されたメモ内の語と発表論文内の語との関連性と，メモ内に主観が反映された語が含まれるのかについて調査した．

5.3.1 コンテキストが反映されたメモとコンテンツ内容との関連性

memoQ では，発表の聴講者が質問メモと個人メモの 2 種類のメモを入力できるように設計していた．当初，質問メモにはユーザが発表内容と関係することを記述すると想定していたため，発表内容と関連のある語が多く含まれるものと想定していた．一方の個人メモは，ユーザが発表に対する意見や感想を記述すると想定していたため，ユーザの主観が反映された語が多く含まれるものと想定していた．そこで，各メモに含まれる語とと発表論文内に含まれる語との関連度を調べることで，2 種類のメモに含まれる語に違いがあるのかについて分析を行った．

メモ内の語と発表論文内に含まれる語の関連度を計算するために，[中山 06] によって提案された Wikipedia マイニングから得られた語彙間の関連度を利用した．Wikipedia マイニングとは，Wikipedia の情報からシソーラスを自動構築する手法のことである．Wikipedia を利用するのは，新しい語や口語といったメモに記述されるような語が多く定義されているからである．また，専門用語も数多く定義されている，という利点もある．シソーラスとして世界的に最も有名な WordNet[*2] では，「OWL」や「RDF」といったセマンティック Web において重要な語は定義されていないが，Wikipedia にはこれらの語は定義されている．

Wikipedia マイニングにおける語間の関連度の求め方は以下のとおりである．Wikipedia におけるすべての Web ページ（記事）の集合を $P = \{p_1, p_2, p_3, ..., p_n\}$ と定義する．ページ $p_i (i \leq i \leq n)$ は，Forward Link と Backward Link の 2 種類のリンクを持つ．p_i の Forward Link は，ページ p_i から別のページへジャンプするリンクの集合であり，$F_{p_i} = \{f_{i1}, f_{i2}, f_{i3}, ..., f_{im}\}$ と定義している．Backward Link は別のページからページ p_i へジャンプするリンクの集合であり，$B_{p_i} = \{b_{i1}, b_{i2}, b_{i3}, ..., b_{im}\}$ と定義している．さらに Wikipedia には，ある記事が参照されたときに，別の記事に対して転送するための機能として，Redirect Link があり，$R_{p_i} = \{r_{i1}, r_{i2}, r_{i3}, ..., r_{im}\}$ と定義している．p_i に関係する語彙の一覧とその関係の強さを求める再帰探索アルゴリズム RE は以下のとおりである．

[*2] http://wordnet.princeton.edu/

Algorithm $RE(p_i, weight, depth)$

if $depth > n$ then return;

$F_{p_i} = GetForwardLinks(p_i)$;

for each $(p_i) \in F_{p_i}$ do

$score = weight/|F_{p_i}|$;

$S_{p_j} = S_{p_j} + score$;

$RE(p_j, score, depth + 1)$;

$B_{p_i} = GetBackwardLinks(p_i)$;

for each $(p_j) \in B_{p_i}$ do

$score = weight/|B_{p_i}|$;

$S_{p_j} = S_{p_j} + score$;

$RE(p_j, score, depth + 1)$;

$R_{p_i} = GetRedirectLinks(p_i)$;

for each $(p_j) \in R_{p_i}$ do

$RE(p_j, weight, depth)$;

本研究では，2007年7月の日本語Wikipediaのデータ（616169語）を用いて，Wikipediaマイニングの処理を行うことによって語彙間の関連度を計算した．まず初めに，メモが付与された発表論文やメモに含まれる語がWikipediaで定義されているのかについて調べた．分析にはメモが付与された発表やメモに含まれる文字列を茶筌を利用して形態素解析した後，名詞と未知語と判断されたもののうち，2文字以上の語を利用した．Wikipediaにメモや発表論文内の語が定義されているかどうかは，発表論文のpdfのテキストやメモのテキストから取得した語がWikipediaに定義されているかどうかで調べた．このとき，語がWikipediaに含まれているかどうかを調べており，意味が同じかどうかまでは考慮していない．表5.1は，メモが付与された発表論文のpdfテキストやメモに含まれる語がWikipediaで定義されているかを調べた表である．質問メモや個人メモに含まれる語数は重複するものを除いている．表5.1によると，質問メモおよび個人メモに含まれる語のうち，半数以上がWikipediaで定義されていることが分かった．一方で，発表論文のpdfテキストに含まれる語は，Wikipediaで定義されていない語のほうが多かった．これは，発表論文のpdfのテキストに含まれる英文概要文も調査対象にしたからかもしれない．英

単語は1582語あり，発表論文のpdfテキストにおける英単語の割合は17%である．ちなみに，日本語のWikipediaにおける英単語の割合は3%（英単語数は20376語）である．Wikipediaは，日本語の単語に比べると圧倒的に英単語が少なく，発表論文のpdfテキストの英単語がWikipediaと一致する確率が低くなることが，発表論文のpdfテキスト内の語がWikipediaで定義されていないことが多い原因の一つであろう．実際に，発表論文のpdfテキスト内の英単語がWikipediaで定義されていない語は1526語あった．これらを考慮すると，発表論文のpdfテキストやメモに含まれる語がWikipediaで定義されている割合は決して高くない．しかし，Wikipediaで定義されていない語には人名や，普通名詞が多いので，Wikipediaのデータを利用してもかまわないものと思われる．

表 5.1: メモが付与された発表論文やメモに含まれる語がWikipediaで定義されている数

	Wikipedia で定義されている語数	Wikipedia で定義されていない語数
メモが付与された発表論文	3672	5538
質問メモ	971	835
個人メモ	571	465

メモと発表ページとの関連度の求め方は，メモに含まれる語が発表論文内に含まれていれば1を，Wikipediaマイニングによって求めた関連語が発表ページ内に含まれていればWikipediaマイニングによって得られた関連度を足していく．このように，関連度はメモに含まれる語が発表論文やWikipediaマイニングによって求めた関連語に含まれていれば足していくため，メモの文字列の長ければ関連度が高くなってしまう可能性がある．そこで質問メモと個人メモの文字列の長さに有意差があるかを調べた．表5.2は，質問メモと個人メモの文字列の長さの中央値や四分位偏差値を示している．アンサリ・ブラドレイ検定を行った結果，$p < 0.05$となり，質問メモの文字列の長さと個人メモの文字列の長さの分散に有意差があるということが分かった．これは，質問メモにおいて，雑談で使われたときは短いメモが多く，議事録として使われたときは長いメモが多かったため，分散が大きくなってしまったからだと思われる．次に，マンホイットニーのU検定を行ったところ，U=638928($n.s.$)となり，質問メモの文字列の長さと個人メモの文字列の長さの中央値の差は有意ではなかった．この結果より，前述した関連度の求め方を用いてもかまわないものと判断した．

表 5.2: 個人メモと質問メモの文字列の長さ（バイト）

	個人メモ	質問メモ
中央値	30	29
四分位偏差値	28	28
メモ数	706	1850

図 5.9 は質問メモと発表ページの関連度のヒストグラムを示しており，図 5.10 は個人メモと発表ページの関連度のヒストグラムを示している．

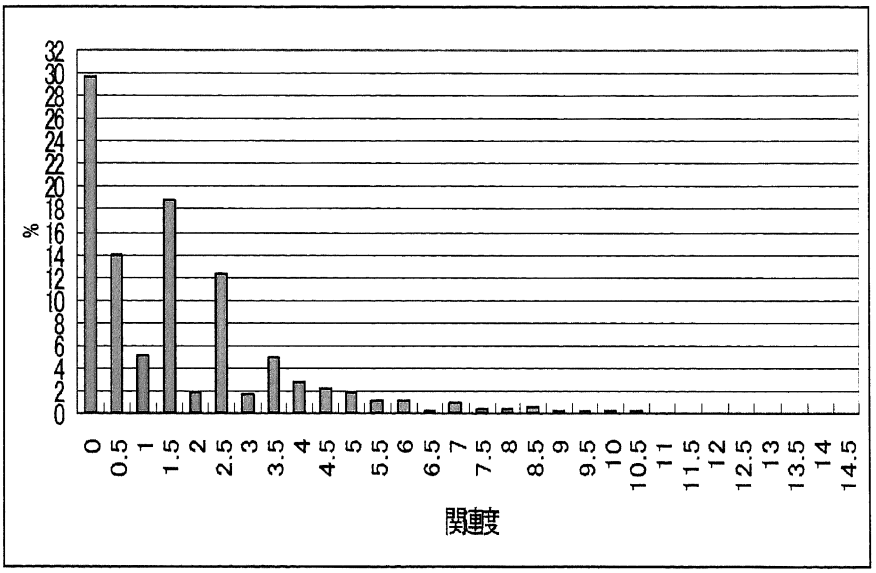

図 5.9: 質問メモと発表論文間の関連度によるヒストグラム

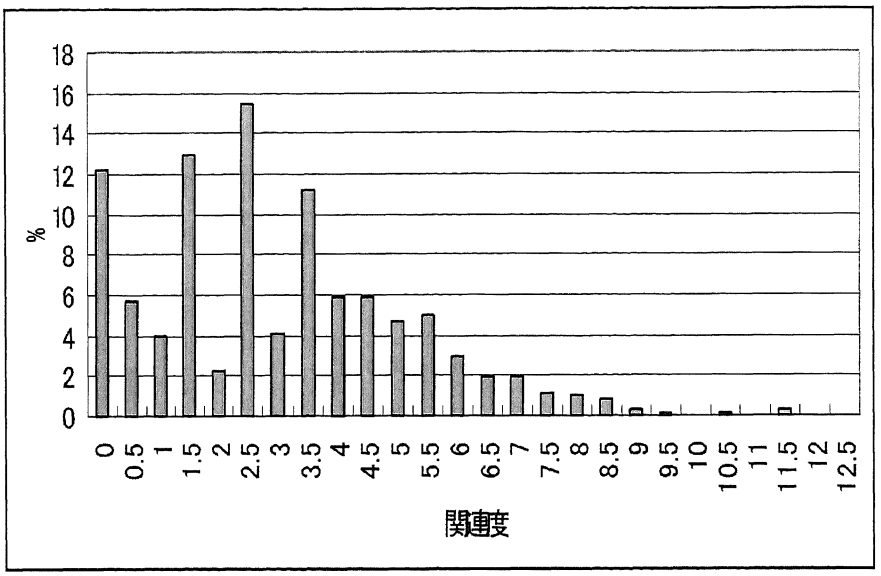

図 5.10: 個人メモと発表論文間の関連度によるヒストグラム

図 5.9 よると，質問メモと発表ページ間の関連度が 0 の割合が高いことが分かる．質問メモには発表内容と関連のある内容が入力されることを想定していたが，実際はそうでないことがわかった．これは，5.2.6 で述べたように，質問メモとして発表内容と関係のないメモが入力されたことが原因の一つだと思われる．一方で，図 5.10 によると，個人メモと発表ページとの関連度が 2 以上の区間の割合が，質問メモの場合の割合よりも高いことが分かる．これは個人メモのほうが質問メモよりも発表ページとの関連性が高いということを示している．表 5.3 は，メモ内の語数とメモ内の語が発表ページに含まれる回数を示している．これによると，個人メモの語が発表論文に含まれる割合が 69.8%（1708/2448 × 100）と，質問メモ内の語が発表論文に含まれる割合の 46.0%（2315/5004 × 100）よりも高いため，ユーザは個人メモをノートととして使っていた可能性がある．

表 5.3: 個人メモと質問メモ内の語が発表論文に含まれる回数と含まれない回数

	メモ内の語が発表論文に含まれる回数	メモ内の語が発表論文に含まれない回数	合計
個人メモ	1708	740	5004
質問メモ	2315	2689	2448

これを調べるために，ユーザが入力したメモ内の語と発表論文内の特徴語との関係について調べた．調査には発表論文内の語の tfidf 値を利用した．メモ内の語の tfidf 値は，メモが付与された発表論文に含まれる語の tfidf 値を計算した後，質問メモおよび個人メモ内の語が発表論文内に含まれていれば，その tfidf 値を利用した．表 5.4 は，質問メモと個人メモ内に含まれる語の tfidf 値の中央値や四分位偏差値を示している．

表 5.4: 個人メモと質問メモに含まれる語に関する tfidf 値

	個人メモ	質問メモ
中央値	1.4	1.0
四分位偏差	5.1	2.3
語数	208	452

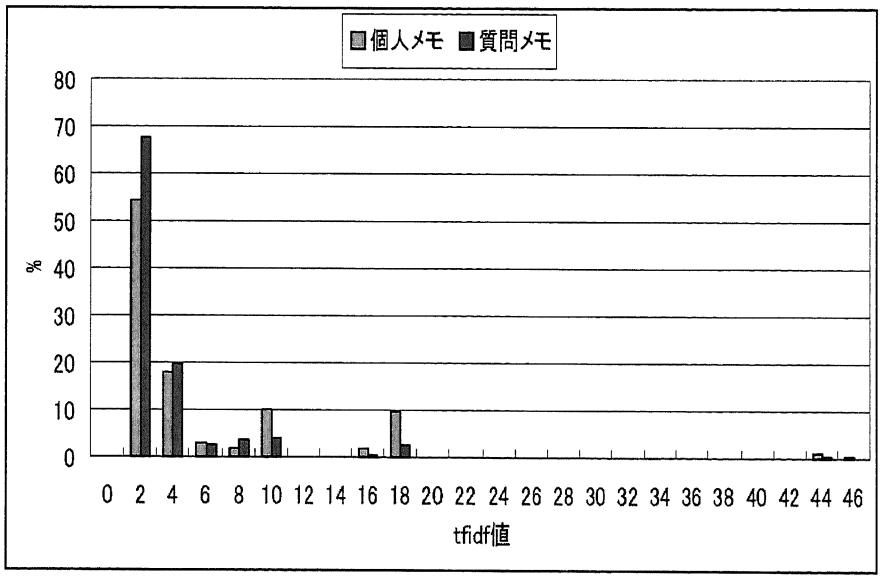

図 5.11: 個人メモと質問メモに含まれる語の tfidf 値のヒストグラム

アンサリ・ブラドレイ検定を行った結果，$p < 0.05$ となり，質問メモ内の語の tfidf 値と個人メモ内の tfidf 値の分散に有意差があるということが分かった．次に，マンホイットニーの U 検定を行ったところ，U=36814.5($p < 0.05$) となり，質問メモ内の語群の tfidf 値と個人メモ内の語群の tfidf 値の中央値の差が有意であった．図 5.11 の個人メモと質問メモに含まれる語の tfidf 値のヒストグラムによると，tfidf 値の 10 以降の区間において個人

メモに含まれる語の出現頻度が質問メモによるものよりも高い．これらの結果より，個人メモには発表論文内の tfidf 値の高い語が多く含まれることから，発表論文内の特徴語が多く含まれることがわかった．したがって，ユーザは発表の内容の要約を記述するために，個人メモを利用していることが示唆された．

実際に，個人メモに書かれている文例は下記のようなものがある．

- 目的：空間データとメタデータ（描画データと呼称）を分離し、後者をパーソナライズすることにより汎用的で有用な地図データを作成する。
- 問題意識：従来の数学書による体系的な理解ではなく、工学で数学を用いるような道具としての学習に向かない。そして一部だけ読んだらそれはそれで理解できない
- 間接投票数の計算など、Google のページランクと類似。人間関係のネットワークを仮定し委任の推移率を行う。
- 新規ユーザ，新規アイテムに対応した協調フィルタリングを用いた推薦システム
- 共益と私益のジレンマ状況での消費者の意思決定の分析

これらは，発表論文と関連度の高い上位 5 つの個人メモである．内容としては，発表における概要を記述したものが多い．このように，個人メモは，発表のノートとして使われる傾向があるため，発表論文との関連度が高いという結果がでたものと思われる．

一方の質問メモは関連度が 0 の割合が非常に高かった．そこで関連度が 0 の質問メモについて調べるために，5.2.6 で利用した質問メモの分類を使って，関連度が 0 の質問メモを分類した．質問メモの分類は，分類 1 が発表に対する質問・感想・意見・ノートで，分類 2 が実際にあった質疑応答のメモ で，分類 3 が発表内容と関係のないメモである． 関連度が 0 の質問メモのうち，1 に分類されるメモは 160 個，2 に分類されるメモは 0 個，3 に分類されるメモは 387 個だった．全質問メモのうち，1 に分類されたメモは 1112 個，2 に分類されたメモは 27 個，3 に分類されたメモは 711 個だったため，3 に分類されたメモが関連度 0 に出現する割合が高いことがわかる．したがって，質問メモの関連度が 0 の割合が高かったのは，発表内容と関係のないメモが多く入力されていたからである．

関連度が 1 より小さい場合，メモ内の語が発表論文内に含まれないことを意味する．そこで，関連度が 1 より小さくて，0 よりも大きい値のメモに発表論文と関連のある語が含まれるかどうかについて調べた．関連度が 0.01〜0.5 までの個人メモおよび質問メモの文例は，下記のようなものがある．

- 口コミを使う.
  タイトル:「オントロジーを用いたWebからの評判情報抽出サービス」

- 文を係り受けの枝ごとに分析する
  タイトル:「構文意味解析における文脈の利用方法」

- XEROX PARCでやられていたColabからの根本的な発展は何なんだろう?
  タイトル:「ディスカッションメディア:会議コンテンツの構造化と効率的な閲覧システム」

- GoogleEarthみたいに昼夜の切れ目が出る?
  タイトル:「オープンコンテント方式にもとづく大規模仮想都市の構築」

- Windowsフォトギャラリー?
  タイトル:「カレンダーを介した新時代のインタラクション」

- ニコニコ動画?
  タイトル:「Synvie: ビデオブログコミュニティから獲得されたアノテーションに基づく応用」

- SNSよりはwikipedia路線か?
  タイトル:「Polyphonet 常時運用の試みテスト運用からの知見と今後の展開」

これらのメモに含まれる語は，発表論文には書かれていないが，発表内容と関連のある用語やソフトウェアやアプリケーションなどである．このように，発表論文には含まれないが，発表内容と関連する語が個人メモおよび質問メモに含まれることがわかった．

5.3.2 コンテキストが反映されたメモと主観が反映された語の関係性

本節では，質問メモおよび個人メモにユーザの主観が反映された語が含まれていたのかについて調べた．はてなブックマークで使用されているタグのうち，下記のタグを主観が反映されている語と判断した．

表5.5で示した語が，質問メモと個人メモに含まれているかどうかについて調査した（表5.6）．なお，発表内容と関係のないメモと判断したメモは調査対象としていない．

表 5.5: はてなブックマークのタグで使用されている主観が反映された語

☆，あとで，あとでみる，あとでよむ，あとで見る，あとで試す，あとで読む，いい話，おもしろ，おもしろい，お役立ち，お笑い，かわいい，これはいい，これはすごい，これはひどい，これは便利，なるほど

表 5.6: 主観が反映された語が含まれるメモ

おもしろい	「具体的に浮かんではいないのですが，コミュニティ支援であるならば，メンバー同士の上下関係なども考慮してもおもしろいかも?」
役立つ	「汎用オントロジーが抽出されたとき，それは概念定義に役立つのか？」，「関連しているのはある程度当然だから、学習に役立つと思うかまで聞くべきでは？」
いい	「にしはらんくいいですよね？」,「こみゅブラいいですね？」
すごい	「144 人中 120 人、すごい回収率？」,「文字検索すごい !?」
便利	「2005 年はべる鈴便利というコメントが多かったです？」,「次世代 Mac のタイムマシーンとどっちが便利か?」,「これは便利、ML の根幹システムに実装してほしい !?! 」
なるほど	「大阪と東京とか複数の相矛盾したものが出現したら？→氏名との近さで判別、なるほど。」

表 5.6 によると，主観が反映された語を含むメモはすべて質問メモによるものだった．個人メモにはこれらの語は含まれていなかった．ユーザは個人の意見や感想を質問メモとして書く傾向があったため，質問メモに主観が反映された語が含まれていたものと思われる．

5.4 考察

本章では，コンテンツ内に含まれないがコンテンツの内容と関連のある語が記述された意味的なメタデータを生成するために，発表聴講時に入力されたメモ内の語を調査した．当初は，質問メモには発表内容と関連のある語が，個人メモにはユーザの主観が反映された語が多く含まれるものと想定していた．5.3.1 の分析結果より，個人メモのほうが質問メモよりも発表論文との関連度が高いことがわかった．また，個人メモのほうが質問メモよりも tfidf 値の高い語が多く含まれていたことから，発表論文内の特徴語をを多く含むことが分かった．これは，ユーザが個人メモを発表聴講時のノートとして利用していたからである．

一方の質問メモは，関連度が 0 の割合が高く，発表内容とまったく関係のないメモが入力

されていたことがわかった．また，5.3.2 の分析結果より，質問メモにだけ主観が反映された語が含まれることがわかった．発表内容とまったく関係のないメモが多く入力されてしまったのは，5.2.9 で述べたように，質問メモが時系列で表示されたため，質問メモに雑談を入力するユーザが増えてしまったからと思われる．質問メモに主観が反映された語が含まれていたのは，ユーザが思いついたことをそのまま入力したからだと思われる．memoQ の設計段階において，ユーザが思いついたことをすばやく入力できるように Twitter の入力インタフェースを採用した．その結果，ユーザは思いついたことを口語で入力したものと思われる．ただし，5.3.2 で示した，主観が反映された語が含まれる質問メモの例を見ると，中には発表内容と関連のある語が含まれているので，質問メモから有用な語を抽出する仕組みが必要であることがわかった．

また，個人メモおよび質問メモの両方において，コンテンツとの関連度が低いメモ内の語をみると，発表論文には含まれないが発表内容と関連のある語が含まれていた．質問メモのほうが個人メモよりも入力された回数が多いため，比較はできないが，質問メモのほうが発表内容と関連のある語が入力されていた回数が多い．したがって，雑談のような発表内容と関係のないメモが入力されないような仕組みを提案すれば，質問メモからは発表聴講によって想起された関連ある語が入力されるようになるのではないかと考えられる．

その他の考察としては，Wikipedia マイニングによって作成したシソーラスを利用して計算したメモとコンテンツ間の関連度が 0 の質問メモには，発表内容と関係のないメモが多く含まれていたことから，発表内容とまったく関係のないメモを取り除く手法として利用できる可能性があるものと思われる．

5.5 まとめ

本章では，2 種類のメモを入力できるシステム memoQ の運用によって得られたデータを使って，メモからコンテンツに含まれないがコンテンツの内容と関連のある語を獲得できるのかについて調査した．memoQ は，ユーザがメモ内に “?” を記述すると質問メモとして他者に公開し，“?” を記述しなければ個人メモとして他者に公開されないような仕組みが提供されていた．分析の結果，個人メモには tfidf 値の高い語が多く含まれており，発表論文内の特徴語が多く含まれることがわかった．一方の質問メモには発表聴講によって想起された質問が入力されていたため，発表論文には記述されていないが，発表内容と関連のある語が獲得できることがわかった．しかし質問メモには，雑談のような発表内容とまったく関係のない語が多く含まれていたので，分析で使用したようなシソーラスを利用

した関連度計算といった手法を用いることによって，発表内容と関連のない語を取り除く必要がある．

第 6 章
結言

本章では，本研究をとおしての成果をまとめ，将来研究の展望について述べて，本論文の結びとする．

6.1 結論

本研究では，WWW 上に存在する多種多様な大量の情報の中からユーザがほしい情報が探しやすくするために，Web コンテンツに意味的メタデータを生成することを目標とした．意味的なメタデータには，Web コンテンツの主題を表す語や特徴語，コンテンツの内容に関連する語，コンテンツ内の語の意味の定義が記述されていることが望ましい．本研究では，意味的なメタデータを生成するために，Web コンテンツの読者である複数ユーザによる下線引きやメモ書きといったアノテーション行為を利用することを提案した．そこで人工知能学会全国大会で運用したアノテーションシステムの運用によって得られたデータを使って，各種分析を行った．

3 章では，ユーザが下線引きによって着目した語にはどのような特徴があるのかについて，JSAI2005 で運用されたイロノミーによって得られたデータを使って分析を行った．その結果，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分かった．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下線が付与された語は，意味的メタデータの生成に利用できる可能性があることが見出された．

4 章では，ユーザがマーキングを付与した箇所を複数ユーザ間で共有した場合，情報探索に役立つのかについて，JSAI2006 で運用した合口によって得られたデータを使って分析を行った．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．また，ユーザはシステムによって推薦されたページよりも，他人がマーキング文字列から他のページへ張ったリンクを選択することも示唆された．さらに，ユーザが付与したマーキング文字列内の語のうち，tfidf 値の高い語が情報探索に有益かどうかについて調べたところ，ユーザは tfidf 値の高い語とは関係なく，他人が付与したマーキング文字列内の語を利用した情報探索を好むということが分かった．これらの結果より，他人がマーキングによってページ内で注目した語は tfidf 値の低い語でも情報探索に有益であることが見出された．

5 章では，ユーザが Web コンテンツに付与するメモからコンテンツに含まれないがコンテンツの内容と関連する語が獲得できるかどうかについて，JSAI2007 で運用した memoQ によって得られたデータを使って分析を行った．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．

イロノミーの分析結果より，ユーザが下線引きによって Web コンテンツ内で注目した語は，コンテンツの内の特徴語や主題を表す語が含まれていたので，コンテンツの要点を示す語として意味的メタデータとして利用できることがわかった．また合口の分析結果より，コンテンツ内において特徴度が低いとされる語でも，同じ興味をもつユーザ同士でマーキング情報を共有すれば，情報を探すときに役立つことがわかった．memoQ の分析結果からは，ユーザがメモを書くときの状況を利用することによって，コンテンツ内の特徴語やコンテンツに含まれないがコンテンツ内容と関連のある語といった異なる種類の語を獲得できることが示唆された．したがって，ユーザがメモを入力するインタフェースを工夫することによって，意味的メタデータに有用な語を獲得できる可能性を見出すことができた．

最後に，今回運用したアノテーションシステムにおいて，集合知が機能していたのかについて述べる．集合知では，集団が多様性，独立性，分散性，集約性という四つの要件を満たしているときに，集団が出す答えの平均値は正解に近く，個々のユーザの答えよりも優れている．イロノミーの分析結果より，個々のユーザの下線の引き方をみると，文書内の特徴度が低い語に下線を付与するものもいたが，ユーザ全体で平均すると下線が付与された語は特徴度の高い語が多いという分析結果が得られた．ユーザが下線を付与するという行為は，他人の意見に影響をうけることが少なく，独立性が実現されていたため，集合知が機能したものと思われる．また，合口の分析結果においても，マーキングによる情報探索が有効だったため，下線引きやマーキングといったアノテーション行為はユーザの集合知を獲得するのに有効な方法といえる．一方，memoQ の分析結果からは，ユーザ間で

時系列でメモを共有した際，他人の意見に影響されて，発表内容とは関係のないメモを入力するユーザがいたことがわかった．[森 06] が指摘しているように，特定の集団内で発言量が多い人の意見にユーザは影響されてしまう，という問題がある．今回は人工知能学会全国大会において運用したアノテーションシステムのデータを収集することによって分析を行った．人工知能学会全国大会は参加者同士が知り合いのことが多く，システムを利用したのも知り合いが多い．memoQ では質問メモを匿名で公開したが，知り合いが多い環境だと，メモの発言内容から誰が入力したのかを推測できる．こうした状況の下，誰かが雑談のような発表内容と関係のないメモを入力すると，他のユーザがそれに反応してしまう，という現象が起きてしまった．したがって，メモを集約するときは他人の影響を受けにくくするような表示をする必要があることがわかった．

これらの結果より，集合知が機能する環境が整っていれば，複数ユーザが Web コンテンツに付与したアノテーションから意味的メタデータを生成するときに有用な語を獲得できる可能性があることが示された．

6.2 課題と今後の展望

本節では，WWW 上の Web コンテンツを対象としたアノテーションシステムを運用するときの課題と，意味的メタデータの流通に向けた今後の将来研究について述べる．

WWW 上の Web コンテンツ全体を対象としたアノテーションシステムを運用する際，解決しなければいけない課題がいくつかある．ネタりか[*1] は，ニュース記事に対して複数ユーザが下線を付与できるシステムである．下線を付与したデータはユーザ間で共有されるため，記事内で複数のユーザが同じ箇所に下線を付与した場合，下線の色が濃く表示されるようになっている．このため，複数ユーザが下線を付与したニュース記事を見た際，下線の色が濃い箇所を見ることによって，注目すべき箇所がすぐ分かるという利点がある．これは，下線を付与する対象がニュース記事のため，下線引きの共有が上手く機能しているのかもしれない．ニュース記事の内容は事象を客観的に捉えたものが多く，また，正しい文法で書かれている．本研究で運用したアノテーションシステムにおいても，アノテーションの対象コンテンツは，ニュース記事と同じ特徴を持っている学会論文だった．このようなコンテンツに対してユーザが下線を引く場合，コンテンツ内容の要点にユーザが注目する可能性が高い．現在，WWW 上にたくさんある Weblog の記事は，一般人が記述することが多いため，ユーザの主観が反映されていたり，間違った文法で書かれていたりす

[*1] http://netallica.yahoo.co.jp/

る．このようなコンテンツをアノテーションの対象とした場合，ユーザが下線引きによって注目する場所が散らばってしまう可能性がある．したがって，下線引きの箇所を共有する際は，下線を付与するコンテンツに注意しなければならない．また，合口の分析結果より，共通の興味を持ったコミュニティ内において，マーキングを共有した場合に情報探索が成功していたことがわかった．したがって，WWW 上でコミュニティ内のユーザとマーキングを共有しない場合は，自分と似た興味を持ったユーザを探す必要がある．

意味的メタデータの流通に向けた将来研究としては，コンテンツ内の語の意味の定義を行うことが挙げられる．RDF や OWL といったメタデータの記述フォーマットや，Swoogle[Ding 04] や Watson[d'Aquin 07] といったメタデータを対象にした検索エンジンといった技術は発展してきているため，意味的メタデータをどうやって生成するかという研究課題は必要不可欠な課題といえる．解決策の一つとして，本研究で提案したような複数ユーザがコンテンツに付与したアノテーションを利用することによって，コンテンツ内の語と既存のオントロジのコンセプトとをマッチングすることが考えられる．たとえば，ユーザがマーキングを付与した語の周辺語や，ユーザが付与したメモ内の語の共起関係を利用することによって，精度の高いコンセプトマッピングが実現するかもしれない．

WWW 上にはネタりかだけでなく，Diigo[*2]，MyStickies[*3]，Fleck[*4] のような商用のアノテーション共有システムが多数存在する．こういったシステムで生成されたアノテーションデータを利用して意味的メタデータを生成することができれば，WWW 上の情報が探しやすくなる環境が整っていくものと思われる．

[*2] http://www.diigo.com/
[*3] http://www.mystickies.com/
[*4] http://www.fleck.com/

謝辞

本研究は，多くの方々のご指導とご助力のもとに遂行されました．以下に特にお世話になった方々のお名前を記して感謝の意を表します．

まず何より，指導教員である国立情報学研究所の武田英明教授に感謝いたします．武田先生には，ご多忙にもかかわらず，いつでも嫌な顔をせずに，丁寧なご指導をしていただきました．こうして本論文を完成させられたのも，ひとえに武田先生のご指導によるものです．ありがとうございました．

本博士論文の審査委員をご快諾くださいました国立情報学研究所の相原健郎準教授，市瀬龍太郎準教授，北本朝展準教授，東京大学の松尾豊准教授にも深く感謝いたします．
相原先生には，研究に対して理解あるコメントをたくさんいただき，大変心強かったです．市瀬龍太郎先生には，常に適確なご助言をいただき，研究を進める上で大変参考になりました．北本先生には，研究成果の主張における詰めの甘さをご指摘いただき，論文の完成度を高めることができました．松尾先生には，イベント支援空間プロジェクトの際にもお世話になりました．

本研究における，3 章および 4 章は，(株) 国際電気通信基礎技術研究所の坂本竜基博士との共同研究です．坂本氏は，JSAI2005 のときに研究に関する相談をしたのをきっかけに，長らくお世話になりました．修士課程時代の研究分野とは異なる分野に飛び込んだ私に，辛抱強く研究指導して下さいました．数々のご助言は，論文を書くときや研究を進めるときの支えになっております．また，(株) 国際電気通信基礎技術研究所 知識科学研究所の小暮潔所長にも大変お世話になりました．ご多忙にもかかわらず，論文に対して数々のコメントをして頂きました．東京農工大学の伊藤禎宣特任准教授および新潟国際情報大学の中田豊久講師は，イロノミーの開発に携わっており，両氏がいなければ私の研究は成り立ちませんでした．大変感謝しております．

学術会議におけるシステムの運用に際しては，イベント空間情報支援プロジェクトのみ

なさまに多大なるご助力を賜りました．

プロジェクトのリーダである産業技術総合研究所の西村拓一博士には，このような機会を賜りましたこと，お礼を申し上げます．また，産業技術総合研究所の中村嘉志博士には，大会支援システムの，主としてインフラの整備においてご尽力いただきました．石田啓介氏には，Polyphonetの開発担当者として，プログラミングやサーバの管理にご尽力いただきました．合口およびmemoQの開発においては，石田氏の技術的支援なくしてはありませんでした．(株) 国際電気通信基礎技術研究所の高橋徹博士はTelMeaシステムを提供されました．藤岡由季氏には，プロジェクトのデザイナとして，Webインタフェースやパンフレット，ポスタのデザインに活躍いただきました．藤村憲之氏には，メディアアーティストとして，研究者や技術者とは異なる視点から，Tabletop Communityの担当者として参加いただきました．皆様，本当にありがとうございました．

本研究は，上記プロジェクト関係者以外にも多数の研究者のみなさまのご指導やご助言を受けて遂行されました．

私が総合研究大学院大学に入学したきっかけは，JAWS2003にて本位田研究室の方々とお会いしたことでした．以後，大学院入学にあたり，東京大学・国立情報学研究所の本位田真一教授を始め，国立情報学研究所の石川冬樹助教，および本位田研究室のメンバーには大変お世話になりました．研究室を移動してからも，食事に誘っていただき，仲良くさせていただきました．感謝しております．

武田研究室のメンバには，公私に渡り多数お世話になりました．国立情報学研究所の大向一輝助教には，総合研究大学院大学における武田研の先輩として，数々のご指導を賜りました．研究室ミーティングでは鋭いご指摘をして頂き，研究の質を高めることができました．産業技術総合研究所の濱崎雅弘博士には，暖かいアドバイスを頂き，いつも励まされました．東京大学の沼晃介博士には，研究に関して数々の助言を頂き，日本電気株式会社の平田敏行博士には，研究に関するご相談をさせていただきました．両氏は，気軽に相談できる先輩として仲良くさせてもらいました．感謝しております．青山大学の鈴木聡博士には，本論文を作成するにあたり，数々の助言をいただきました．こうして完成できたのも氏のおかげです．ありがとうございました．総合研究大学院大学の後藤孝行氏には，memoQの開発にあたり，大変お世話になりました．インタフェースの研究分野についての知見は参考になりました．その他，東京大学の福原知宏博士を始め，総合研究大学院大学の小出誠二氏，荒木次郎氏，(株) アルファシステムズの丹英之氏，慶應義塾大学の深見

嘉明氏，東京大学の亀田尭宙氏，研究室秘書の澤田幸氏をはじめとする新旧のメンバに感謝いたします．

またこのほか，国立情報学研究所において研究生活をともに送ってきた総合研究大学院大学や連携大学院の同僚や先輩方にも感謝いたします．

冒頭にも述べましたとおり，本研究は多くの方々のご助力によって成すことができました．ここに記したのはその一部であり，お名前を記すことのできなかった多くのみなさまにも，感謝いたします．

最後に，家族への感謝をもって結びます．父 秀輔と母 恭子には，長い学生生活を支えていただきました．今後もがんばっていきたい所存です．

みなさま，本当にありがとうございました．

参考文献

[Baca 98] Baca, M. and Gilliland-Swetland, A.: *Introduction to Metadata: Pathways to Digital Information*, Getty Publications (1998)

[Berners-Lee 01] Berners-Lee, T., Handler, J., and Lassila, O.: The Semantic Web, *Scientific American* (2001)

[Blanchard 87] Blanchard, J. and Mikkelson, V.: Underlining Performance Outcomes in Expository Text, *Journal of Educational Research*, Vol. 80, No. 4, pp. 197–201 (1987)

[Brickley 04] Brickley, D.: RDF Vocabulary Description Language 1.0: RDF Schema, `http://www.w3.org/TR/rdf-schema/` (2004)

[Collier 04] Collier, N., Kawazoe, A., Kitamoto, A., Wattarujeekrit, T., Mizuta, Y., , and Mullen, A.: Integrating Deep and Shallow Semantic Structures in Open Ontology Forge, 人工知能学会 セマンティックウェブとオントロジー研究会, 第 SIG-SWO-A402-05 巻 (2004), (in English)

[Damme 07] Damme, C. V., Hepp, M., and Siorpaes, K.: FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies, in *Bridging the Gep between Semantic Web and Web 2.0 (SemNet 2007)*, pp. 57–70 (2007)

[d'Aquin 07] d'Aquin, M., Gridinoc, L., Sabou, M., Angeletou, S., and Motta, E.: Watson: Supporting Next Generation Semantic Web Applications, in *WWW/Internet conference 2007* (2007)

[Dave Raggett 99] Dave Raggett, A. L. H. and Jacobs, I.: HTML 4.01 Specification, `http://www.w3.org/TR/html4/` (1999)

[Davis 95] Davis, J. R. and Huttenlocher, D. P.: Shared annotation for cooperative learning, in *CSCL '95: The first international conference on Computer support for*

collaborative learning, pp. 84–88, Mahwah, NJ, USA (1995), Lawrence Erlbaum Associates, Inc.

[Dawson 98] Dawson, F. and Howes, T.: vCard MIME Directory Profile, `http://tools.ietf.org/html/rfc2426` (1998)

[Denoue 00] Denoue, L. and Vignollet, L.: An annotation tool for web browsers and its applications to information retrieval (2000)

[Dill 03] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A., and Zien, J. Y.: SemTag and seeker: bootstrapping the semantic web via automated semantic annotation, in *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pp. 178–186, New York, NY, USA (2003), ACM

[Ding 04] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., and Sachs, J.: Swoogle: a search and metadata engine for the semantic web, in *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp. 652–659, New York, NY, USA (2004), ACM

[Golder 06] Golder, S. A. and Huberman, B. A.: Usage patterns of collaborative tagging systems, in *Journal of Information Science*, pp. 98–208 (2006)

[Gyunn 78] Gyunn, S. M.: Capturing reader's attention by means of typographical cueing strategies, Vol. 18, pp. 7–12 (1978)

[Handschuh 03] Handschuh, S., Staab, S., and Volz, R.: On deep annotation, in *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pp. 431–438, New York, NY, USA (2003), ACM

[Joachims 05] Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G.: Accurately Interpreting Clickthrough Data as Implicit Feedback, in *Proceedings of the Conference on Research and Development in Information Retrieval* (2005)

[Joinson 99] Joinson, A.: Social desirability, anonymity, and Internet-based questionnaires, and Internet-based questionnaires, in *Instruments and Computers*, pp. 433–438 (1999)

[Kahan 01] Kahan, J., Koivunen, M.-R., Hommeaux, E. P., and Swick, R. R.: Annotea: an open RDF infrastructure for shared Web annotations, in *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pp. 623–632, New York, NY, USA (2001), ACM

[Kashyap 97] Kashyap, V. and Shet, A.: Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies, in *Cooperative Information Systems*, pp. 139–178 (1997)

[Khare 06] Khare, R.: Microformats: The Next (Small) Thing on the Semantic Web?, *IEEE Internet Computing*, Vol. 10, No. 1, pp. 68–75 (2006)

[Kolari 06] Kolari, P., Finin, T., and Josh, A.: SVMs for the Blogosphere: Blog Identification and Splog Detection, in *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs* (2006)

[Lawrence 00] Lawrence, S. and Giles, C. L.: Accessibility of information on the Web, *Intelligence*, Vol. 11, No. 1, pp. 32–39 (2000)

[Marshall 97] Marshall, C. C.: Annotation: from paper books to the digital library, in *DL '97: Proceedings of the second ACM international conference on Digital libraries*, pp. 131–140, New York, NY, USA (1997), ACM

[Mathes 04] Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Metadata, `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html` (2004)

[Matthews 83] Matthews, J. R., Lawrence, G. S., and Ferguson, D.: *Using Online Catalogs: A Nationwide Survey*, Neal-Schuman Publishers, Inc., New York, NY, USA (1983)

[McBride 04] McBride, B.: Resource Description Framework (RDF): Concepts and Abstract Syntax, `http://www.w3.org/TR/rdf-concepts/` (2004)

[McGuinness 04] McGuinness, D. L. and Harmelen, van F.: OWL Web Ontology Language Overview, `http://www.w3.org/TR/owl-features/` (2004)

[Meyer 01] Meyer, E. A. and Bos, B.: Introduction to CSS3, `http://www.w3.org/TR/css3-roadmap/` (2001)

[Mika 05] Mika, P.: Ontologies are us: A Unified Model of Social Networks and Semantics, in *Proceedings of the Fourth International Semantic Web Conference (ISWC2005)*, pp. 5–15 (2005)

[Morris 77] Morris, C. D., Bransford, J. D., and Franks, J. J.: Levels of Processing versus Transfer Appropriate Processing, *Journal of Verbal Learning and Verbal Behavior*, Vol. 16, No. 5, pp. 519–533 (1977)

[O'Reilly 05] O'Reilly, T.: What Is Web 2.0 : Design Patterns and Business Models for the Next Generation of Software, `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html` (2005)

[Pemberton 02] Pemberton, S., et al.: XHTML? 1.0 The Extensible HyperText Markup Language (Second Edition), `http://www.w3.org/TR/xhtml1/` (2002)

[Pilgrim 02] Pilgrim, M.: What is RSS, O'Reilly XML.com (2002)

[Popov 03] Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M.: Towards Semantic Web Information Extraction, in *Human Language Technologies Workshop at the 2nd International Semantic Web Conference (ISWC2003)*, Florida, USA (2003)

[Resnick 94] Resnick, P., Lacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, in *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pp. 175–186 (1994)

[Röscheisen 94] Röscheisen, M., Mogensen, C., and Winograd, T.: Shared Web Annotations As A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples, Technical report, Computer Science Department, Stanford University (1994)

[Salton 91] Salton, G.: Developments in automatic text retrieval, *Science*, Vol. 253, pp. 974–979 (1991)

[Specia 07] Specia, L. and Motta, E.: Integrating Folksonomies with the Semantic Web, in *Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*, pp. 624–639 (2007)

[Steve DeRose 01] Steve DeRose, E. M. and Jr., R. D.: XML Pointer Language (XPointer) Version 1.0, `http://www.w3.org/TR/WD-xptr` (2001)

[Stuckenschmidt 04] Stuckenschmidt, H. and Harmelen, van F.: *Information Sharing on the Semantic Web*, Springer (2004)

[Surowiecki 05] Surowiecki, J.: *The Wisdom of Crowds*, Anchor, Garden City (2005)

[武田 04] 武田 英明, 大向 一輝 : Weblog の現在と展望—セマンティック Web とソーシャルネットワーキングの基盤として—, 情報処理, Vol. 45, No. 6, pp. 635–661 (2004)

[Tim Berners-Lee 05] Tim Berners-Lee, R. F. and Masinter, L.: Uniform Resource Identifier (URI): Generic Syntax, `http://tools.ietf.org/html/rfc3986` (2005)

[Vargas-Vera 02] Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., and Ciravegna, F.: MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup, in *EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pp. 379–391, London, UK (2002), Springer-Verlag

[Xian Wu 06] Xian Wu, L. Z. and Yu, Y.: Exploring social annotations for the semantic web, in *Proceedings of the 15th international conference on World Wide Web (WWW2006)*, pp. 417–426 (2006)

[Yee 02] Yee, K.-P.: CritLink: Advanced Hyperlinks Enable Public Annotation on the Web (2002)

[坂本 06] 坂本 竜基, 中田 豊久, 伊藤 禎宣, 松岡 有希, 小暮 潔, 武田 英明 : イロノミー : 色付き傍線による Web 文章を対象としたフォークソノミー, 第 20 回人工知能学会全国大会 (JSAI2006) 論文集 (2006)

[松岡 07] 松岡 有希, 坂本 竜基, 伊藤 禎宣, 武田 英明, 小暮 潔：選択文字列を用いた Web ページ推薦システムでのユーザ参加型リンクアンカ付与機能の実証実験による評価, 第 21 回人工知能学会全国大会 (JSAI2007) 論文集 (2007)

[松本 03] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸：日本語形態素解析システム「茶筌」Version2.3.3, http://chasen.naist.jp/hiki/ChaSen/ (2003)

[沼 07] 沼 晃介, 平田 敏之, 濱崎 雅弘, 大向 一輝, 市瀬 龍太郎, 武田 英明：学術会議における体験共有のための行動履歴に基づく Weblog システム, 情報処理学会論文誌, Vol. 48, No. 1 (2007)

[森 06] 森 健：グーグル・アマゾン化する社会, 光文社 (2006)

[西村 04] 西村 拓一, 濱崎 雅弘, 松尾 豊, 大向 一輝, 友部 博教, 武田 英明：2003 年度人工知能学会全国大会支援統合システム, 人工知能学会誌, Vol. 19, No. 1, pp. 43–51 (2004)

[西田 06] 西田 健志, 五十嵐 健夫：Lock-on-Chat: 複数の話題に分散した会話を促進するチャットシステム, 日本ソフトウェア科学会論文誌 コンピュータソフトウェア, Vol. 23, No. 4, pp. 69–75 (2006)

[中山 06] 中山 浩太郎, 原 隆浩, 章治郎 西尾：Wikipedia マイニングによるシソーラス辞書の構築手法, 情報処理学会論文誌, Vol. 47, No. 10, pp. 2917–2928 (2006)

[土方 04] 土方 嘉徳：情報推薦・情報フィルタリングのためのユーザプロファイリング技術, 人工知能学会誌, Vol. 19, No. 3, pp. 365–372 (2004)

[日本 04] 日本図書館情報学会研究委員会：図書館目録とメタデータ-情報の組織化における新たな可能性, 勉誠出版 (2004)

[武田 06] 武田 英明, 松尾 豊, 濱崎 雅弘, 沼 晃介, 中村 嘉志, 西村 拓一：イベント空間におけるコミュニケーション支援, 電子情報通信学会誌, Vol. 89, No. 3, pp. 206–212 (2006)

[齋藤 03] 齋藤 孝：三色ボールペン情報活用術, 角川書店 (2003)

研究業績

査読付き学会誌論文

1. 松岡 有希, 坂本 竜基, 伊藤 禎宣, 大向 一輝, 武田 英明, 小暮 潔: **マーキングを用いたソーシャルタギングの有効性に関する検証**, 情報処理学会論文誌, Vol. 48, No. 12, 2007

査読付き国際会議等発表

1. Yuki Matsuoka, Ikki Ohmukai and Hideaki Takeda: **COLLABORATIVE MEMO-BASED SYSTEM IN THE ACADEMIC CONFERENCE**, In *Proceedings of IADIS International Conference WWW/Internet 2007 (ICWI2007)*, 2007.

2. Yuki Matsuoka, Ikki Ohmukai and Hideaki Takeda: **Working Towards Ontology Generation from Context of Listening to Presentations**, In *Poster Proceedings of 6th International Semantic Web Conference (ISWC2007)*, 2007.

3. Yuki Matsuoka, Ryuuki Sakamoto, Sasanori Ito, Hideaki Takeda and Kiyoshi Kogure: **Aikuchi: Marking-based Social Navigation System**, In *Poster Proceedings of International Conference on Weblogs and Social Media 2007 (ICWSM2007)*, 2007.

4. Yuki Matsuoka, Ryuuki Sakamoto, Sasanori Ito, Hideaki Takeda and Kiyoshi Kogure: **Social tagging using marked strings in web pages**, In *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW2006) located at the 5th International Semantic Web Conference (ISWC 2006)*, 2006.

査読付き国内学会発表

1. 松岡 有希, 坂本 竜基, 中田 豊久, 伊藤 禎宣, 武田 英明: **論文概要に対する色付きアンダーライン付きステムの運用・分析**, 電子情報通信学会 第17回データ工学ワークショップ (DEWS2006) 論文集, 2006.

査読なし学会等発表

1. 松岡 有希, 武田英明 : **アノテーションを用いた学会発表聴講支援システムの提案**, 第8回 AI 若手の集い（MYCOM2007）, 2007.

2. 松岡 有希, 坂本 竜基, 伊藤 禎宣, 武田 英明, 小暮 潔: **選択文字列を用いた Web ページ推薦システムでのユーザ参加型リンクアンカ付与機能の実証実験による評価**, 人工知能学会全国大会 (第21回) 論文集, 2007

3. 松岡 有希, 坂本 竜基, 伊藤 禎宣, 武田 英明, 小暮 潔: **Web 文書に対するマーキングからの個人知識の獲得**, 人工知能学会全国大会 (第20回) 論文集, 2006

付 録 A

システム実装

合口と memoQ の実装について述べる.

A.1 合口の実装

図 A.1 は，合口の実装図である.

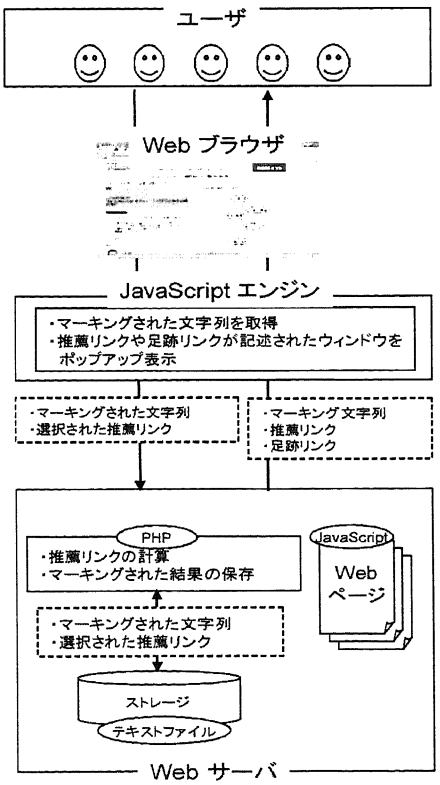

図 A.1: 合口の実装図

合口は，ユーザクライアント側とWebサーバ側で実行される．それぞれ，JavascriptとPHPで実装した．ユーザがWebページにアクセスしたときに，Javascriptエンジンは Webサーバにマーキング文字列があるかどうかを問い合わせる．ユーザがWebブラウザ上で文字列をマウスカーソルでなぞったとき，Javascriptエンジンはユーザ IDとマーキングされた文字列の情報をWebサーバに送る．Webサーバは受け取ったユーザ IDとマーキングされた文字列を使って，推薦リンクを計算する．Webサーバが推薦リンクの計算結果をJavascriptエンジンに送ると，Javascriptエンジンは推薦リンクを記述したポップアップウィンドウをユーザに表示する．もしユーザが推薦リンクの一つを選んだら，Javascriptエンジンは Webサーバにその情報を送り，テキストファイルに下記の情報をテキストファイルに保存する．

- 日付
- ユーザ ID
- マーキングされた文字列
- Webページ上でマーキングされた文字列の位置
- 選択された推薦リンクの URL
- 選択された推薦リンクの計算に用いた推薦アルゴリズム

ユーザがマウスカーソルをマーキング文字列の上をなぞると，Javascriptエンジンはマーキング文字列の情報をWebサーバに送る．Webサーバはマーキング文字列の情報を受け取って推薦リンクを計算し，ストレージから足跡リンクの情報を取得する．Webサーバが推薦リンクと足跡リンクをJavascriptエンジンに送ると，Javascriptエンジンは推薦リンクと足跡リンクを記述したポップアップウィンドウをユーザに表示する．ユーザが推薦リンクか足跡リンクのどれか一つを選ぶと，Javascriptエンジンは Webサーバにその情報を送り，Webサーバはテキストファイルに保存する．

A.2 memoQの実装

図 A.2は，memoQの実装図である．

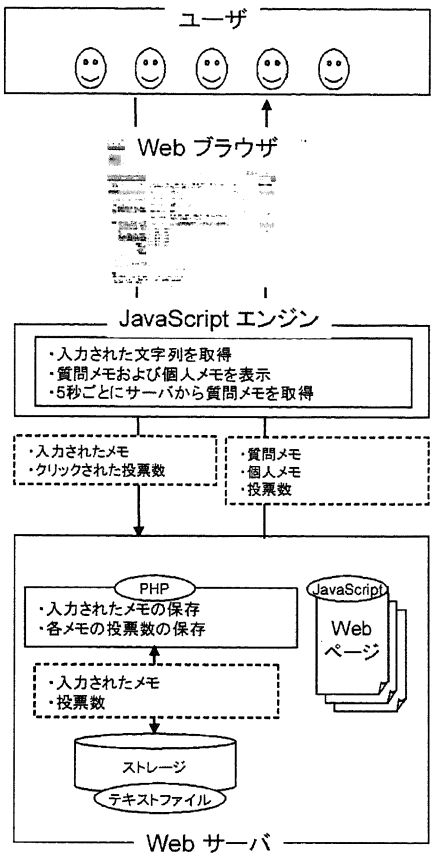

図 A.2: memoQ の実装図

memoQ も合口と同様に，サーバ側で PHP を，クライアント側で Javascript を用いて実装した．ユーザがメモを入力すると，Japascript エンジンは Web サーバにメモの文字列を送ると同時にメモに?が含まれていれば質問メモとして，含まれていなければ個人メモとして表示した．Web サーバはメモの情報を受け取ると，テキストファイルに下記の情報をテキストファイルに保存する．

- 日付
- ユーザ ID
- 発表ページ ID
- メモ ID

- 入力されたメモの文字列
- メモを入力したユーザの投票数（初期値は，個人メモは 0，質問メモは?の数)
- 全ユーザの投票数

また，ユーザが質問メモに投票するために，ラジオボタンをクリックすると，JavaScript エンジンは投票数をサーバに送り，サーバは投票数をテキストファイルに保存する．他人の質問メモを表示するために，JavaScript エンジンは，5 秒ごとにサーバ側に質問メモと投票数を取得するようにした．

付録 B

memoQ運用後のアンケートとその結果

第21回人工知能学会終了後に，memoQに関するアンケートを行った．回答者は27名だった．

Q1. memoQは発表の聴講時に役立ちましたか？

大変役に立った	5
役に立った	13
どちらとも言えない	4
あまり役に立たなかった	3
まったく役に立たなかった	2

Q2. 質問メモは匿名で公開されるので，入力しやすかったですか？

入力しやすかった	12
どちらとも言えない	14
入力しにくかった	1

Q3. 他の方が入力した質問メモをリアルタイムに見ることは，ご自身にとってどういう影響がありましたか？

- 他人の質問に突っ込みたくなってしまった．
- 他の人がどのように感じながら発表を聞いているのか知ることができ、疑問の共有や解決に役に立った。
- 臨場感があった
- 時間切れで質問できなかったときに見たので……
- 自分にはなかった視点に気づけた。
- 色々なポイントの気づきをリアルタイムに助けてくれたが、聴講への注意がそがれた
- 同じ質問を持っている人などがいることが分かり良かった．
- なんとなく盛り上がりがわかる。

- 面白いがプレゼンターへの意識は薄れる
- 発表に対する聴講者の本音や関心の有無を知ることができて，ためになった． ただ，匿名のためか，あまり参考にならない書き込みも多く見受けられた．
- 質問メモをみることによって新しい質問を思いつくという創発効果があった．
- 面白かった．
- 多くの方から意見を頂けてよかった
- 返事をしたいが返事の出し方がわからなかった。
- 他の人の視点でも発表を見ることができる点がよかった．
- ちゃんとした質問は参考になったが、ただのチャットになっているときには、聴講に集中できなくなった。
- 学会の経験が浅い僕にとって、発表を聞く視点を与えてくれたという側面もあった。
- 発表者にとっては、発表時の質疑応答とは別に会場の人たちの本音が聞けるために、厳しい意見もありますが、今後のためにはとても有益な意見となりました。
- 質問されない方のその時その時の考えも分かるので良いと思う。
- リアルタイムでは使っていませんでした．
- 考えることは皆，似たりよったりなので，memoQ で質問をまとめなくても，同じ種類の質問が得られるかなと思いました．
- 自分と同じ質問があるので入力の手間が省ける

Q4 質問メモのまとめ図は役に立ちましたか？

大変役に立った	1
役に立った	6
どちらとも言えない	3
あまり役に立たなかった	7
まったく役に立たなかった	2
そんな機能があることを知らなかった	6

Q5. ご意見・ご感想があれば教えてください.

- 質問の状況を発表者に見えるようにするか，利用者同士の議論もできるようにしておいてもらえないとつまらない． 発表の向上のためにも，もっと自由なシステムの方が良かったのでは
- チャットとしての利用だったので、本来の使い方であるメモとして役に立ったかとなると少し評価が落ちる。他人の質問に投票できる機能は非常によかったと思う。
- メモを消す機能もほしかった
- ?を入れると公開になるが，返事をするにも？を付けないといけない． 返事用の方法も必要だと思う． 匿名なのはよいが，自分の発言だけは自分に分からないと，あとで返事をしてくれたかどうか見てまわれない．
- みんなが投票すると色が濃くなるのが面白く、みんなが投票してくれるような質問を考えていた。

- あまり適切でない発言もあったので、マイナス評価ボタンも欲しかった。
- 座長をさせていただいたとき，memoQ を使うことで会場からの質問を吸い上げることができてよかった．ただ，まだ利用者があまり多くなかったのが残念でした．
- ノート PC を持参して聴講しないと使えないのが最大のネックと思う
- 質問と回答をセットで表示できる方がわかりやすいと思った．全発表の質問と回答を一覧できるような機能や，どの発表に対して，いつ質問・回答が行われたかをリアルタイムで知ることができると良いと思った（RSS のようなもので）．また，その際に自分の研究分野と関連のある発表をフィルタリングできるとさらにうれしい．発表中の質疑応答を録音しておいて，(半) 自動的に memoQ に反映させることは可能でしょうか？
- 荒れたときもあったのが残念
- 匿名にすると，無責任な発言が多くなるので，できれば匿名でない方がいいと思いました．
- 発表後、発表者が memoA するというのはあっていいかなと思う。ActionLog の方に memoA してた人がいたけど、memoQ と連携してそういう機能があればいいと思う。自分の質問が赤くなったり取り上げられたときは嬉しかった。それ以外の要望は polyphonet の memoQ コミュに投稿しています。
- 難しいとは思いますが、誹謗中傷につながるコメントがなるべく少なくなるようにするとよいと思います。
- 各発表のトップに最新のメモ Q が表示されると良かった気がします．(もしかして表示されてました？) メモ Q の存在自体を忘れることが多かったです．
- 「？」をつけたものだけが公開されることを知らなかったので，積極的に利用しませんでした．メモした時間が管理できるのは便利で良いと思います．今年は使い方を把握していなかったのですが，来年は活用したいと思いました．できれば，発表者のスライドに関連付けてメモできると便利だと思いましたが，さすがに難しそうですね．
- このシステム自体は大変良いと思いますが、匿名で好きなことが書き込めることから、2ch と同様と感じている参加者がいたようで、2ch を見ているような書き込みが 連続したこともありました。例えば、発表者、質問者の発言に対して、「○○キター」など。学生さんではないかと思いますが、研究内容の質問、疑問ではなく、個人的な上辺の部分を思ったままに書き込みしているように感じ、学会会場でこのような書き込みがあることにとても幻滅しました。参加者の学会への参加意識が薄いのではないでしょうか。
- ２ちゃんねる化をどう防ぐかが課題でしょうか。
- Mac および新しい Windows でも動作しなかった。同じ会場内で、利用できる人とそうでない人に分かれると、利用できない人に疎外感が生まれると思った。聴講している人みんなの考えや、その場の雰囲気が分かることには意味があると思った。