アノテーション共有システムにおける投稿の有用性判断方式

佐伯　唯†　遠山　元道††

† 慶應義塾大学大学院理工学研究科　〒 223–8522 神奈川県横浜市港北区日吉
†† 慶應義塾大学理工学部情報工学科　〒 223–8522 神奈川県横浜市港北区日吉
E-mail: †saeki@db.ics.keio.ac.jp, ††toyama@ics.keio.ac.jp

あらまし　アノテーション共有システムとは，限定されたグループ内でキーワードに登録したアノテーションを共有するツールであり，任意の Web ページでキーワードにマウスオーバーすることで全てのアノテーションを閲覧することができる．これはグループでの共同研究・作業における情報や認識の共有を容易にすることを目的としているが，任意のページでの閲覧を可能にしていることから，必要のないアノテーションが表示されることがある．それを解決するために，有用性の高い投稿のみを表示するための有用性判断方式を提案する．有用性判断にはアノテーションを投稿した際の Web ページのトピックと共有時の Web ページのトピックの近さを算出し，有用性を判断する．

キーワード　アノテーション，共同作業コンピューター支援，有用性判断

1　はじめに

近年，インターネットの普及により情報共有は E メールや SNS，ドライブ等を用いて主に行なわれている．しかし，会社内や研究室における共同作業・研究において一つ一つの情報や認識を即時に共有するためにはそれぞれの情報を自らの手で所定の場所に保存する，もしくは送信する必要があり非常に手間がかかる．Web ページ内において何か共有したい情報が出てきた場合，その場でその単語に直接全員に共有できるメモを書き込むことができれば共同作業・研究がより効率的になるのではないかと考えた．そこで，著者らの以前の研究 [1] において特定のグループ内で，Web ページ内の単語に直接アノテーションを付ける形でその内容を共有する関数型アノテーション共有システムを提案した．

関数型アノテーション共有システムを実装するために，著者らが提案，開発を行なっている Web IndeX システム [2] [3] を応用した．Web IndeX(WIX) システムとは，Web ページの閲覧者が閲覧中のページ内に登場する単語に関連する他の Web ページへのアクセスを容易にするために，キーワードと URL の組み合わせであるエントリを XML 形式で記述した WIX ファイルを用いて，Web 文書中のキーワードからハイパーリンクを生成するシステムである．

既存のツールに多い特定の Web ページのみにアノテーションを登録するシステムを付箋型とし，それに対して任意のページ内の単語でアノテーションを共有できる本システムを関数型と位置付けた．

関数型システムは，特定のページで付加されたアノテーションを任意のページで閲覧することができるため，アノテーションを効率的に閲覧することができる．しかし，その Web ページを閲覧する際に必要のないアノテーションや，信頼性の低いアノテーションが表示されることがある．例えば，図 1 では他のページで付けられたハイブリッド車に関するアノテーション

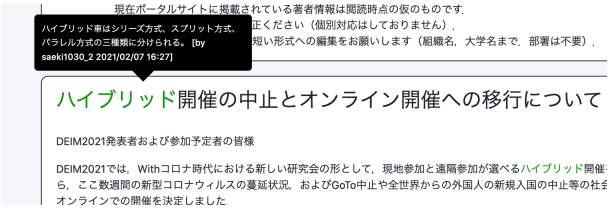

図 1　必要のないアノテーション例

が，関係の無い DEIM ハイブリッド開催に関わる記事 [4] で表示されている．これを解決するために，本研究では有用性判断の基準を検討し，アノテーションのランク付けを行い上位数件のみを共有時に表示する手法を提案する．

本論文の構成は以下の通りである．まず 2 章においてアノテーション共有システムについて述べる．3 章でアノテーション共有システムの課題，4 章で有用性判断のための提案手法，5 章では関連研究について述べる．6 章で評価をし，7 章で今後の課題について述べる．

2　アノテーション共有システム

2.1　概　　要

本論文で使用するアノテーション共有システムは、Chrome 拡張機能を用いて実装した．システムの概要を図 2 に示した．ユーザーはブラウザ上のメニュー画面からデータベースへユーザー登録及びアノテーションの投稿を行い，サーバー側で辞書式マッチングのための Find Index [5] を構築する．ユーザーがブラウザ上で共有内容を閲覧する際には，共有エンジンが閲覧中の Web ページの文章を書き換える．

ブラウザ右上のボタンで表示される図 3 のポップアップメニューからグループ登録，アノテーション登録を行い，ブラウザ下部の図 4 のツールバーから参加グループを選択し共有内容を表示する．

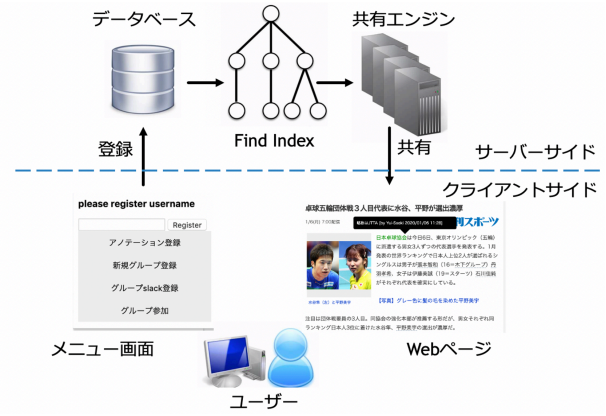

図 2　アノテーション共有システムの概要

図 3　ポップアップメニュー

図 4　ツールバー

2.2　ユーザー・グループ管理

ユーザー管理には，chrome 内のデータストレージである chrome storage を使用した．オプションで chrome アカウントと同期する機能を選択し，異なる端末においても同じ chrome アカウントであれば同一のユーザーとして自動的にログインできる機能を実装した．一つの chrome アカウントにつき初回のみユーザー名登録が必要であるため，システム導入時にのみ登録画面が表示され，それ以後はメニュー画面が毎回表示される．ユーザー名を登録すると，データベースのにユーザー名が登録される．

また，グループはグループ名とパスワードによって管理されており，ユーザーはメニュー画面から新たなグループを登録することができ，登録されたパスワードを入力することでグループへの参加ができる．

2.3　アノテーション登録

アノテーションを登録する方法は，ポップアップメニューからの登録とブラウザ上で単語を選択し右クリックでの登録の 2 種類実装した．ポップアップメニューからの登録では図 5 のように単語をユーザーが入力し，右クリックによる登録では図 6 のように選択した単語が登録画面で表示される．アノテーションを登録すると，DB 内にキーワード，アノテーション内容が格納される．

図 5　メニューからの登録

図 6　右クリックからの登録

2.4　使　用　例

Web IndeX システムと同様に，ツールバーのボタンを押すことで Web ページの変換が行なわれる．変換後の Web ページの例を図 7 に示す．データベース内で選択したグループにおいて登録されている単語が全て緑色に変化し，その単語にマウスオーバーすることで黒い吹き出しでアノテーション内容が表示される．図 7 の例には日刊スポーツのウェブニュース記事 [12] を使用した．

図 7　共　有　後

3　アノテーション共有システムの課題

3.1　関数型と付箋型の比較

前章で述べたアノテーション共有システムは，任意のページにおける単語そのものにアノテーションを登録する関数型である．関数型と従来のアノテーション共有システムである付箋型でアノテーションにどのような特徴があるか比較を行った結果を表 1 に示した．

表 1　関数型と付箋型の比較

	長所	短所
付箋型	意図を正確に伝えられる	閲覧する機会が少ない
関数型	効率的に共有できる	有用ではない投稿が表示される

関数型の短所である有用ではない投稿が表示されるという点に着目し，実際に関数型アノテーション共有システムを使用する場合に有用性のないものが表示されるかを確かめるために次節で予備実験を行った．

3.2　予 備 実 験

予備実験では，研究室内の 10 人が実際にアノテーションを投稿するという作業を 2 週間に渡って行った．同じ言葉に対するアノテーション比較を容易にするため，対象の web ページは指定したニュース記事 2 件に絞って実験を行った．

この実験から，以下のことが分かった．

- 抽象的、または感情的な内容のアノテーションは他ページで閲覧しても意味がない．
- どのページで元々付けられたアノテーションなのかを知りたい
- 別ページの URL を含んだ投稿が多い

ここでの「抽象的，または感情的な内容」というのは，難しい，面白いといった対象に対する言及が無いアノテーションを指す．

この実験により，関数型アノテーション共有システムにおいて考えられる問題点が 3 点判明した．一つ目は，アノテーションとして別ページの URL が含まれる場合，そのページは信頼できるものであるかという点．二つ目は，長い時間が経過したアノテーションは最近のものより有用性が劣るという点．最後にアノテーションとそれが付けられた言葉に関連性が薄いと他ページ閲覧の際の有用性が下がる点である．

4　提 案 手 法

4.1　判 断 方 式

前節で判明した問題点を踏まえ，複数の観点からアノテーションの投稿を評価し有用性判断を行い，判断結果をランク付けした上位数件のみを共有時に表示するという方式を提案する．また，アノテーションのみでは有用性が低いと考えられる感情表現などを抽出し，アノテーションと共に元のページの情報を付加する方式も検討している．提案方式の概要を図 8 に示す．

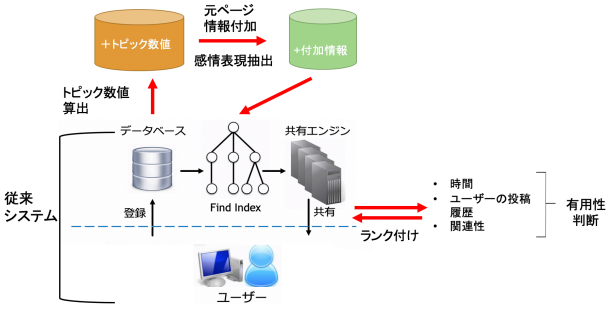

図 8　有用性判断方式の概要

この方式で用いる有用性判断の基準は以下の 4 点である．

- 投稿から閲覧までの期間
- アノテーションと付加されたワード，閲覧中のページ内容との関連性
- 投稿に他ページ URL が含まれる場合，そのページの信頼性
- 有用性が高い投稿が多いユーザーであるか

4.2　提案手法の概要

判断基準 2 点目である，アノテーションと付加されたワード，閲覧中のページ内容との関連性を判断する手法を提案する．

関連性を判断するために，それぞれの内容の話題の近さの測定を行う．話題を数値として取得するために，トピックモデルの手法である LDA [6] を利用する．この時のトピック数は固定しておらず，トピック数による結果の違いを評価する予定である．今回は単語の出現頻度からのアプローチである LDA を使用したが，単語の分散表現からのアプローチである Word2vec や文章の分散表現からのアプローチである Doc2vec を利用した手法も実装する．

また，問題点で言及した抽象的，感情的な内容のアノテーションの場合には，共有時にアノテーション投稿時の Web ページの概要を追加で掲載する手法も検討している．その際にはアノテーション内容から感情を表す言葉を自然言語処理を用いて抽出する予定である．

4.3　提案手法のシステムへの組み込み

まずはアノテーション投稿時に，使用された Web ページのトピックを算出する．そのトピックの数値をアノテーションの内容と共にデータベースに格納する．

アノテーションを共有する際には，閲覧している Web ページのトピックとその Web ページに出現している各ワードのトピックの近さを算出し，各アノテーションで近いものから有用性が高いと判断し上位に表示する．

5　関 連 研 究

5.1　アノテーション共有システム

本研究で使用したシステムと同様にアノテーションを共有するシステムとして，Addison Y.S. Su らが開発した PAMS2.0 [7] が挙げられる．このシステムは主に教育用に開発されたもので

あり，限定された学生のグループ内で同一の資料を読みながらアノテーションを書き込み共有することで理解を深めることを目的としている．PAMS2.0 においては，個別でのチャットやグループ全員でのディスカッションを行うことができる．また，アノテーションに質問，解答等の分類分けのタグを付けたり，ページ上に直接文字を書き込んだりすることが可能である．

また，同様に教育目的のシステムとして，Yu-Chien Chen らが開発した MyNote [8] が挙げられる．このシステムは，e ラーニングの学習管理システム (LMS) に組み込まれているシステムであり，LMS 内の学習オブジェクトにアノテーションをつけることが出来る他，LMS から Web ドキュメントにアノテーションをつけることも可能である．

上記の研究が特定のページにアノテーションを付加する付箋型であるのに対し，本研究におけるシステムは任意のページにおける単語そのものにアノテーションを登録する関数型である．

5.2 有用性判断

不特定多数のユーザーが投稿した内容を評価するという点において本研究と関連する研究に，Twitter の投稿におけるフェイクニュース検出と EC サイトでのレビュー評価が挙げられる．

Siva Charan Reddy Gangireddy らの研究 [9] ではフェイクニュースを発信するユーザーの行動を仮定し，グラフベースの教師無し学習を用いている．Qiang Zhang らの研究 [10] では投稿とその返信の内容と時系列を考慮し，フェイクニュース検出のためのベイズ深層学習モデルを提案している．

Debanjan Paul らの研究 [11] では従来の EC サイトのレビュー評価において他ユーザーの評価を用いていたため，他ユーザーの評価が少ないレビューを正しく評価できないという点に着目し動的畳み込みニューラルネットワークを用いたレビュー評価を提案している．また，同様の側面に対してレビューをする場合に別の言葉が使用されることへ自然言語処理を用いて対処した．

6 評　　価

予備実験時と同様に実際に何人かに一定の期間でシステムを使用してもらいアノテーションを収集し，システム内で有用性判断を行った結果をユーザーが有用だと感じた結果と比較することで評価を行う予定である．

今回の関連性判断の手法として提案した LDA を利用した場合と，その他の Word2vec や Doc2vec を利用した場合の関連性判断結果を比較し評価を行う．また，トピックモデルにおけるトピック数や各判断基準を総合してランク付けする際の重み付けなどの変数を変更し，それによる結果の違いも評価する．

7 今後の課題

現在は判断基準の 2 つ目であるアノテーションと付加されたワード，閲覧中のページ内容との関連性を判断する手法のみ実装中であるため，有用性判断の精度を高めるために他 3 点の判断基準もシステムに実装することが今後の課題である．また，それぞれの判断基準で算出したものをどのような重み付けをしてランク付けをするかという点も同時に考えていく．

1 つ目の投稿から閲覧までの期間による有用性判断は，投稿が行われた日時と閲覧時の日時を取得し，その差分を算出するシステムを実装する予定である．また，3,4 点目の他ページの信頼性，投稿ユーザーの信頼度による有用性判断に関しては今後判断手法を検討していく．

文　　献

[1] 佐伯唯，遠山元道．Web IndeX を応用した関数型アノテーション共有システムの実装．第 12 回データ工学と情報マネジメントに関するフォーラム，DEIM2020．2020.

[2] 林昌弘，青山峻，朱成敏，遠山元道. Keio WIX システム (1) ユーザインターフェース. データ工学ワークショップ，DEIM2011. 2011.

[3] 森良介，藪達也，朱成敏，遠山元道. Keio WIX システム (2) サーバーサイド実装. データ工学ワークショップ，DEIM2011. 2011.

[4] DEIM2021 実行委員会『DEIM2021 第 13 回データ工学と情報マネジメントに関するフォーラム』閲覧日時 2021 年 2 月 7 日 https://db-event.jpn.org/deim2021/index.html

[5] 石崎文規，遠山元道. 大規模 Aho-Corasick オートマトンにおける追加更新手法の提案. データ工学ワークショップ，DEIM2012. 2012.

[6] D.M.Blei, A.Y.Ng and M.I.Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, vol. 3, 993-1-22, 2003.

[7] Addison Y.S. Su, Stephen J.H. Yang, Wu-Yuin Hwang b, Jia Zhang. A Web 2.0-based collaborative annotation system for enhancing knowledge sharing in collaborative learning environments. Computers & Education 55, 752-766, 2010.

[8] Yu-Chien Chen, Ren-Hung Hwang, Cheng-Yu Wang. Development and evaluation of a Web 2.0 annotation system as a learning tool in an e-learning environment. Computers & Education 58 (2012), 1094-1105, 2010.

[9] Siva Charan Reddy Gangireddy, Deepak P, Cheng Long Nanyan and Tanmoy Chakraborty. Unsupervised Fake News Detection: A Graph-based Approach. In Proceedings of the 31st ACM Conference on Hypertext and Social Media. ACM, 75-83, 2020.

[10] Qiang Zhang, Aldo Lipani, Shangsong Liang and Emine Yilmaz. Reply-Aided Detection of Misinformation via Bayesian Deep Learning. In Proceedings of the 2019 World Wide Web Conference (WWW ’19). ACM, 2333-2343, 2019.

[11] Debanjan Paul, Sudeshna Sarkar, Muthusamy Chelliah. Recommendation of High Quality Representative Reviews in e-commerce. In Proceedings of the Eleventh ACM Conference on Recommender Systems. ACM, 311-315, 2017.

[12] 日刊スポーツ,『卓球五輪団体戦 3 人目代表に水谷、平野が選出濃厚』閲覧日時 2020 年 1 月 14 日 https://www.nikkansports.com/sports/news/202001050000683.html