編集距離を利用した短文情報推薦システム

津守　優[†]　　大西　克彦[†]　　南角　茂樹[†]

本研究では，文法構造などを考慮しない大量の短文情報からユーザが興味を持つ情報を推薦する手法として，編集距離を利用した情報推薦システムを提案する．提案システムでは，ユーザが興味を持つ短文情報に対して，これらを共有する Web サイトからユーザが興味を持つ内容と類似する情報を推薦する．短文情報では品詞のみの情報などが多く，従来の形態素解析を利用した手法では類似度計算が難しい．そのため本システムでは編集距離を利用して文字列の編集操作を基にした類似度算出手法を利用する．短文情報として Twitter を利用したシステムを実装し，ユーザによる評価の結果，提案システムが推薦した情報について有用性が確認できた．

A Recommendation System for Short Term Information with the Levenshtein Distance

YUTAKA TSUMORI[†]　　KATSUHIKO ONISHI[†]　　SHIGEKI NANKAKU[†]

We proposed a recommendation system which can select the short term information belong to users interest. Our system can recommend a short term information similar to what the user is interested in sharing information from the web site. Because the short term information often have not grammatical structure, it is difficult to calculate similarity by using morphological analysis which is ordinary used in this area. Therefore our system utilize an analysis method by using Levenshtein distance. We evaluate the usefulness of our system which is used Twitter as the short term sharing web site.

1. はじめに

Web などを通じて世の中に流通する文字情報は 10 年前と比較し，約 200%もの伸び率となっている[1]．内容についても一般的な文章から，Twitter などの即時性に着目し文字数を制限した短文情報を発信するサービスなどさまざまなものが利用されている．このような状況の中 Web の世界ではユーザの求める情報に類似する付加情報を提供する機能が普及してきている．これらの機能の多くは形態素解析を利用した手法が多く，類似精度の高い情報が提供できるが，文法構造などをあまり考慮しない短文情報についてはこれまであまり利用されていない．

そこで本研究では，編集距離を利用して大量の短文情報からユーザに推薦できる情報を提示するシステムを検討した．具体的には，まず本研究で利用する編集距離を利用した短文情報での類似情報提示手法の有用性について，従来の形態素解析による手法と比較した．そして短文情報を共有できる Web サービスである Twitter のお気に入り情報を利用し，ユーザが求める情報に類似する情報を推薦できるシステムを提案・実装した．

本稿では第 1 章で概要を述べる．第 2 章では情報推薦システムの従来研究と本提案システムの位置づけについて述べる．第 3 章では本手法で利用する編集距離による推薦手法について述べる．第 4 章では本システムの実装と評価について述べる．

2. 関連研究

ユーザの興味に基づく情報を提示する手法としては，楽天などの広告コンテンツや，Amazon のおすすめ商品などがある[2]．これらの機能はユーザの購買履歴を情報源とし，購買履歴が似通っている他のユーザが購入しているものを紹介する仕組みである．また，Google の広告サービスなどでは現在表示している Web ページの内容を解釈し，内容に関連した広告を自動的に表示する．

このような類似した文字情報の選択には主に対象となる文字情報に対して形態素解析を元にした n-gram 言語モデルを利用して類似度の判定を行っている[3]．しかし，形態素解析や n-gram 言語モデルを利用した手法では形態素解析のための辞書や，n-gram 言語モデルのデータベースなどのリソースが必要となる．

また編集距離を利用した文書情報の類似度を判定する手法として，編集距離と形態素解析を用いた手法によりレポートの剽窃を判定する手法[4]がある．さら

† 大阪電気通信大学　大学院総合情報学研究科
Graduate School of Information Science and Arts, Osaka Electro-Communication University

にコンピュータネットワークを流れるメールの文字情報に対して編集距離を計算しスパムメールの判定に利用する手法が提案されている[5]．しかし，これらの研究では対象となる文書情報はある程度の長さのものであり，短文情報だけを対象にはしていない．

短文情報を対象とした情報の推薦システムについてもこれまでさまざまな手法が提案されている．例えば「ソーシャルニュース記事推薦システム」[6]では，Twitter 内の短文データを TF-IDF 手法を用いて類似度を判定し関連するニュースを提示する．TF-IDF は主に単語の出現頻度に着目し，あらかじめ大量のデータを収集しておく必要がある．また，短文情報からニュース文章などの長文情報を推薦するものであり，短文情報間での類似の情報を提示する手法とは異なる．このように，対象となる文書情報が，長文や文法規則に基づいた日本語で記述された文書であれば，形態素解析や n-gram 言語モデルを利用した類似度算出手法がよく利用されている．しかし短文情報は文法規則に沿った日本語では無い場合が多く，形態素解析による文章の類似度計測が難しい．

そこで，本研究では短文の場合は形態素解析よりも，単純な評価方法として編集距離を利用して文字列の類似度を計測し，ユーザが求める情報に類似したものを推薦できる可能性があると考えた．そこで編集距離を利用した手法と，一般的な形態素解析を利用した手法を用いて短文情報における類似度の違いについて予備実験を行った． 次節でその結果について述べる．

3. 編集距離を利用した推薦手法

3.1 編集距離を利用した推薦

編集距離（レーベンシュタイン距離）は特定の 2 組の文字列に対してその相違を編集操作の回数を利用した指標で示すことができる[8]．つまり，それぞれ異なる文字列 A と文字列 B に対して，文字列 A を文字列 B に編集する際の文字の削除操作と挿入操作の回数を文字列間の一致度の指標として利用する．**図 1** に例を示す．**図 1** では文字列“今日は雨です”から文字列“今日は晴れです”へ編集する場合の手順を示し，1 文字（“雨”）の削除操作と 2 文字（“晴れ”）の挿入操作が含まれるので編集距離は 3 となる．

この編集距離を任意の文字列に対しての類似度の指標として利用できる．例えば．文字列“西名阪自動車道”に対して，文字列 “名阪国道”と文字列“名神高速道路”の類似度を比較する場合，“西名阪自動車道”と“名阪国道”の編集距離は 5 になり，“西名阪自動車道”と“名神高速道路”の編集距離は 9 となることから，文字列“名阪国道”は文字列“名神高速道路”より，文字列“西名阪自動車道”に編集距離が近く類似性が高いといえる．

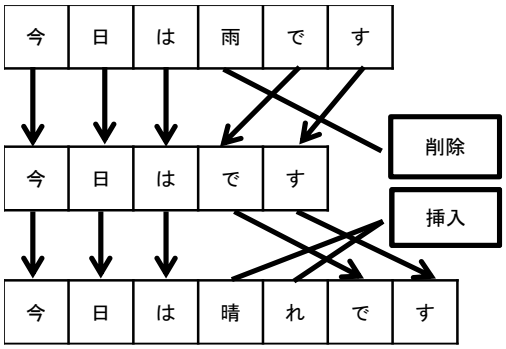


図1 編集距離の計算手順

しかし，これらの文字列の類似度を形態素解析と n-gram で実施した場合，“西名阪自動車道”“名阪国道”“名神高速道路”は形態素解析によりすべて固有名詞として解析され 1 文節になる．そのため， n-gram による類似度を求めると一致度が 0 となり，これらの文字列に対しての類似性を判断できない．

そこで，本研究では固有名詞などの品詞のみで利用されることの多い短文情報について，編集距離を利用した推薦手法について検討する．

3.2 短文情報の推薦手法

本研究で利用する編集距離による推薦手法のアルゴリズムを図 2 に示す．まず，処理する対象の文字列を s1 とし，その文字長を l1 とする．次に最短の編集距離を格納する変数 l2 を十分に大きな値で初期化しておく．その後比較するすべての文字列に対して，s1 との編集距離を計算し最小値を l2 とする．そして，新たに受信した情報の文字列の長さを L，編集距離を L’としたとき，式(1)を満足する場合にユーザに有益な情報としてマーキングを行う．

$$L' \leq L \times \alpha \qquad (1)$$

3.3 短文情報における従来手法との比較

編集距離を利用した類似度による推薦手法について，従来手法として形態素解析を利用した手法と計算速度，精度について比較した予備実験を行った．実験タスクは，ある任意の文字列を推薦元のデータとし，別の短文情報となる文字列に対して類似性の高い文字列を推薦文として抽出する．実験に利用するデータは，Twitter のパブリックタイムラインをクロールし収集した．推薦元のデータは 528 件とし，比較するデータは 32,265 件とした．これらのサンプルデータに含まれる文字列に関連性はなく，Twitter 上で共有されて

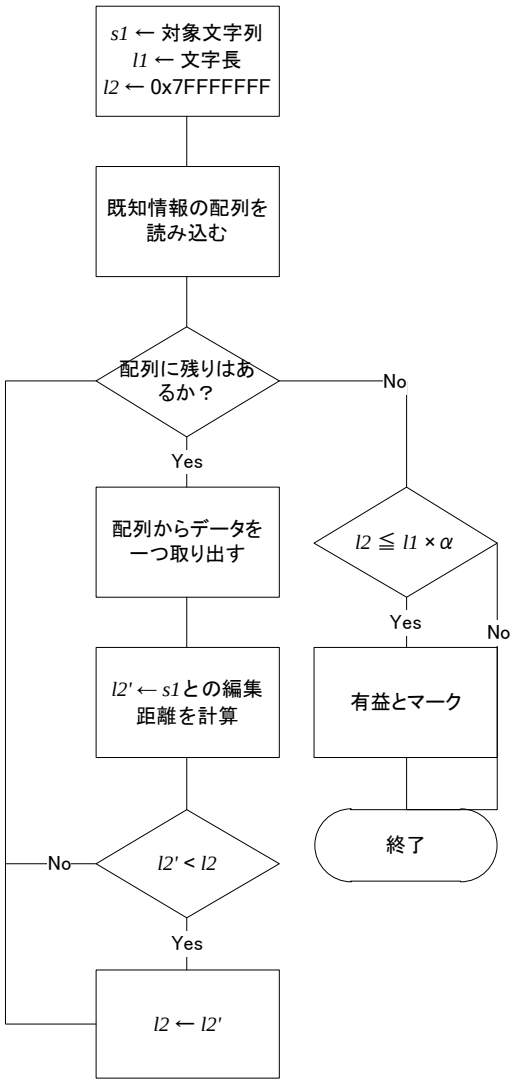


図2 推薦処理アルゴリズム

いる文字列からランダムに抽出した．これらのデータを利用して，編集距離での類似度算出と，形態素解析を用いて品詞分解を行いその結果から 2-gram を用いて類似度を算出した場合について，それぞれの推薦文として抽出された数と処理時間を計測した．なお，編集距離，形態素解析それぞれを利用した手法の推薦基準は，編集距離は前節で述べた手法を用いている．従来手法では，形態素解析によって得られた結果を隣接する 2 つの文節ごとに連結を行い，その組み合わせの個数を元に評価を行った．形態素解析アルゴリズムには MeCab[8]を利用した．実験環境は次の通りである．結果を**表 1** に示す．

- OS: Windows 7 Ultimate x64
- CPU: Intel Core i7 640-LM
- RAM: 8GB

表1 2 手法の評価

手法	推薦された短文数	割合	実行時間
編集距離	1495	4.63%	225.422 秒
形態素解析	820	2.54%	17380.383 秒

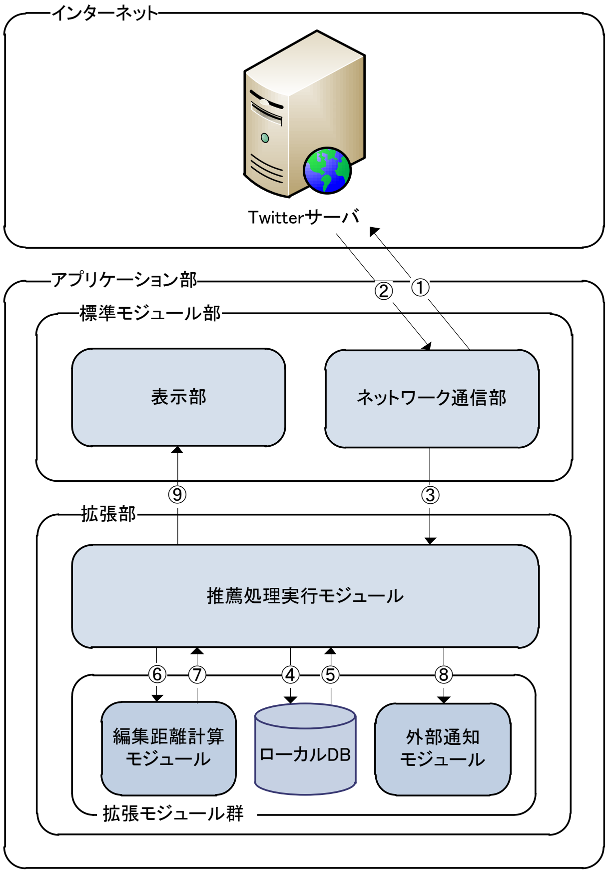
表 1 から，形態素解析と 2-gram を用いた手法の実行時間は，編集距離を用いた場合の約 77 倍となっている．従って，実行時間効率に対する点では編集距離による短文情報推薦手法は一定の効果があると考えられる．

4. 短文情報推薦システム

4.1 システム概要

編集距離による短文情報推薦システムとして，ソーシャルサービスであり，マイクロブログの Twitter を利用する．Twitter は 140 文字までの短文を Web 上で共有し，ユーザ間での情報交換や情報共有ができるサービスである．Twitter には「お気に入り」と呼ばれる機能があり，ユーザが任意の発言に対してマーキングできる．この機能によってマーキングされる発言は，例えば，あとで読みたい URL が含まれている発言や，メモとして保存する発言，発信者に同意する発言などユーザにとって何らかの有益な情報が多く見受けられる．そこで，ユーザのお気に入りの履歴を利用してタイムラインに含まれるステータスに対して推薦を行うシステムを作成する．

本システム全体の概要を**図 3** に示す．最初に Twitter サーバから「タイムライン」と呼ばれる短文情報を取得する．タイムラインは API を使用し一度に 40 件取得するか，または Twitter の Web サーバからプッシュ配信によってネットワーク通信部で取得する．取得した情報は推薦処理実行モジュールへと受け渡される．推薦処理実行モジュールでは，アプリケーションが持つローカルのデータベースよりユーザの「お気に入り」データを読み取る．そして，Twitter サーバから取得したタイムラインの 1 つのデータに対して，ユーザの「お気に入り」データのいくつかを組みにして，編集距離計算モジュールへ受け渡しその編集距離を算出する．この編集距離計算モジュールより返されたすべての編集距離のうち最小になる組みを記録し，その組みの編集距離が閾値以下であれば推薦するデータと定義し，ディスプレイ表示部からユーザへ推薦結果を提示する．この一連の推薦処理では，お気に入りデータを推薦元のデータとして利用するため，お気に入りデータの数が多いユーザに対しては高頻度



① タイムラインデータを要求
② タイムラインデータを受信
③ 受信したデータを推薦処理部へ受け渡し
④ 既知のお気に入りデータを要求
⑤ お気に入りデータを読み取り
⑥ 既知のお気に入りデータと
タイムラインのデータを元に計算を実行
⑦ 計算結果を取得・推薦処理を実行
⑧ 推薦結果を外部通知モジュールへ表示
⑨ 推薦結果を本体モジュールへ返却

図3 今回提案したシステムの概要図

に推薦が行われ，逆にお気に入りデータが少ないユーザに対しては推薦されるデータが少なくなり，ユーザの好みを強く反映する結果になると考えられる．また，本システムは推薦処理を行う際に，受信した短文情報と発信者が同じお気に入りデータのみを使用している．こうすることで，ユーザがお気に入りをたくさん付けるユーザは推薦頻度が高く，お気に入りを付けないユーザの推薦頻度は低くなる．つまり，ユーザのユーザに対する好みも推薦処理の評価基準として用いることが出来る． システムの実装には Web 上で公開されているアプリケーション Azurea を用いた[9]．Azurea は

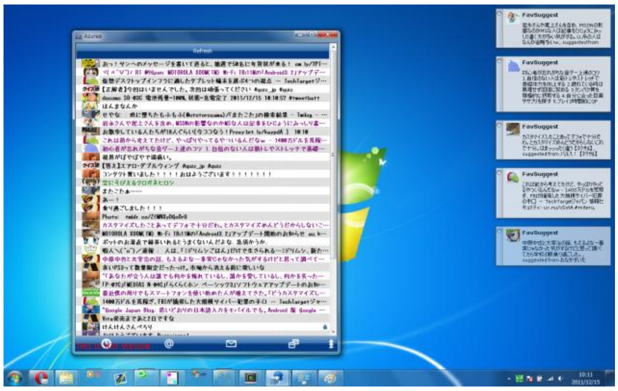


図4 全体図

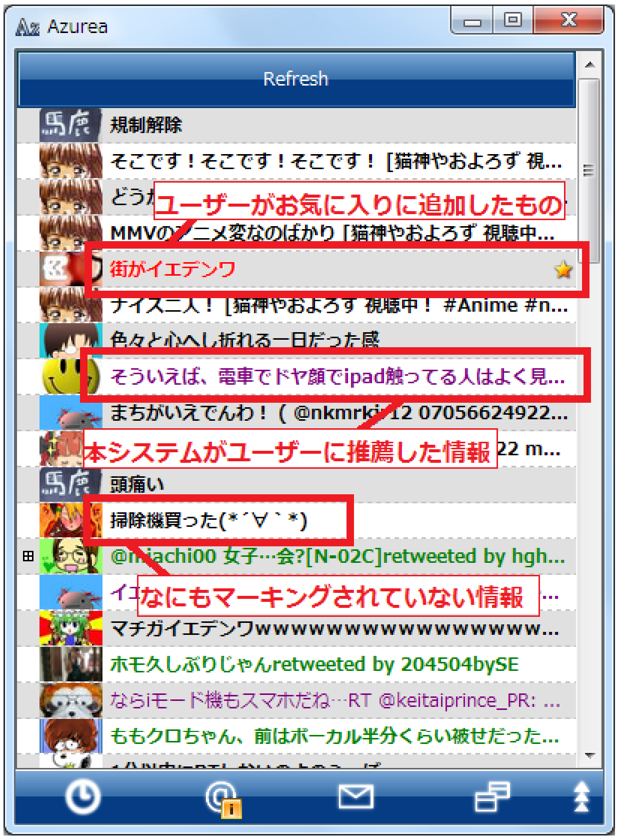


図5 タイムライン表示部

JavaScript や C++などの言語でユーザがある程度機能拡張できる特徴をもっている．本システムの実行画面を**図 4** に示す．また，**図 5** に推薦結果をユーザのタイムライン上に表示した画面を示す．推薦された短文情報は紫色で描画される．本システムでは Twitter で良く利用される文字列の先頭が@から始まり半角英数字および半角_（アンダースコア）が連続する短文の場合，この@から始まる一連の半角文字列は文字列の一致度という観点ではノイズと考えこれらの文字列は削除して推薦処理を実行する．このように推薦された短文情報を提供することで，推薦された短文情報に対して確認し発言者や時間などの情報を一覧できると考

えられる．

4.2　ユーザによる評価と考察

本システムによって推薦される短文情報について実際にユーザに使用してもらい，お気に入り数，推薦された短文情報数，そのうちお気に入りに追加された数を計測し，その感想を聴取した．ユーザは本システムの短文情報推薦機能について説明しシステムの操作を学習し，その後，短文情報の推薦結果を画面上に表示する機能を省いたシステムを利用して，お気に入りを追加した．短文情報の推薦結果はシステム内部で保存した．なお，本システムで利用する編集距離による推薦基準の係数は 0.8 とした．

普段からお気に入り機能をよく利用しているユーザ 5 人に対して本システムを利用してもらい，各ユーザのお気に入り数，推薦された短文情報数，そのうちお気に入りに追加された数を**表 2** にまとめた．また，お気に入り数に対する推薦された短文情報数と，お気に入りに追加した情報数のうち，推薦された短文情報数を比率で表示したグラフを**図 6** に示す．これらの図から，ユーザがお気に入りに追加した短文情報のうち約 25%は本手法によって推薦された短文情報である．また本手法により推薦された短文情報のうち，平均で約 33%の情報がユーザの「お気に入り」に新たに追加されている．この結果から推薦された短文情報には有用性があることが確認できる．

また，実際に使用したユーザからは，自分のお気に入りを学習してくれている印象を受けたことや，自分の嗜好の傾向が反映されていることが明らかに判断できたという意見があった．実際に推薦された例を挙げると“おやつにみたらし団子食べた”に対し“お寺と隣接してるあの茶店のみたらし団子とぜんざい食べたい”や，“ふぁぼられたの？”に対し“最近ふぁぼられまくってるんだけどこれメリットってあるの？”というものが推薦されていた．

表2　お気に入りに追加された数

ユーザ	お気に入り数	推薦数	内お気に入り追加数
ユーザ 1	907	245	80
ユーザ 2	228	86	21
ユーザ 3	607	170	64
ユーザ 4	240	125	15
ユーザ 5	512	1073	48

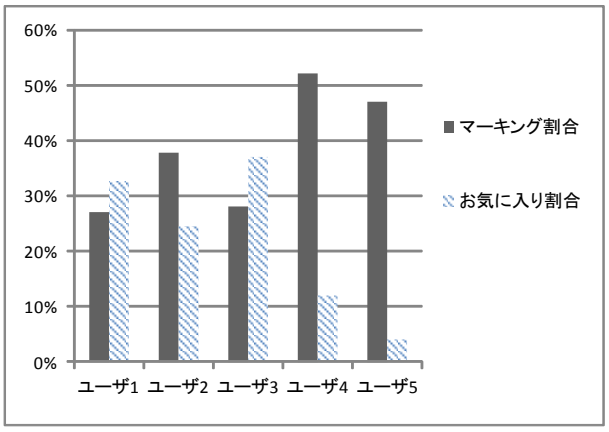


図6　マーキングした情報の割合

5. まとめ

本研究では，文法構造を考慮しない短文情報に対して編集距離を利用した類似度算出手法を利用した情報推薦システムを提案した．短文情報データとして，Twitter のタイムライン上のデータを利用して，ユーザのお気に入りに類似する短文情報を提示するシステムを作成した．ユーザによる評価の結果，お気に入りに追加するデータとの相関があり一定の有用性を確認できた．今後の課題としては，短文情報の推薦手法で利用している類似度算出アルゴリズムの精度向上や推薦情報の提示手法の検討などが挙げられる．

参　考　文　献

1) 情報通信政策研究所調査研究部: 我が国の情報通信市場の実態と情報流通量の計量に関する調査研究結果, http://www.soumu.go.jp/main_content/000124276.pdf , (2011)
2) Greg Linden, Brent Smith, and Jeremy York: Amazon.com recommendations: item-to-item collaborative filtering, Internet Computing, IEEE, Volume 7, Issue 1, pp.76-80 (2003)
3) 天野　真家，宇津呂　武仁，成田　真澄，福本淳一，石崎　俊: 自然言語処理，オーム社 (2007)
4) 上田　和志，富永　浩之: 類似性に基づくレポート剽窃の検出ツールの開発，情報処理学会研究報告．コンピュータと教育研究会報告 2010-CE-107(9), pp.1-8 (2010)
5) 塚島　洋介，中尾　彰宏: 編集距離の最近傍探索を用いたインネットワーク・スパムフィルタリング手法，電子情報通信学会技術研究報告．NS，ネットワークシステム 110(448), pp.449-454, (2011)

6) 早川　豪，岡部　誠，尾内　理紀夫: Twitter を利用したソーシャルニュース記事推薦システム, 情報処理学会研究報告. データベース・システム研究会報告, 2011-DBS-153(16), pp.1-4 (2011)
7) Vladimir Iosifovich Levenshtein: Binary codes capable of coorectiong deletions, insertions, and reversals, Soviet Physics Doklady, Volume 10, Issue 8, pp.707-710 (1966)
8) MeCab: Yet Another Part-of-Speech and Morphological Analyzer: http://mecab.sourceforge.net/
9) Azurea - Twitter client for Windows Mobile devices: http://azurea.info/ja/