マイクロブログ文書の選択による対話的な災害情報検索システム

北口 沙也香 †　　宮西 大樹 ‡　　関 和広 ‡　　上原 邦昭 ‡
† 神戸大学 工学部　‡ 神戸大学大学院 システム情報学研究科
kitaguchi@ai.cs.kobe-u.ac.jp

1　はじめに

東日本大震災発生時，マイクロブログの代表的なサービスである Twitter では情報伝達を目的として震災に関する大量のマイクロブログ文書（ツイート）が投稿された[1]．投稿されたツイートは，現実世界の出来事を即時に反映しており，安否情報や被災者への支援情報といった重要な情報が含まれていた．しかし，震災当時には大量のツイートが氾濫したため，過去と現在の情報や多種多様な情報が入り混じることになった．そのため，大量のツイートの中から自身の意図にあった情報を見つけることは困難を極めた[2]．そこで，本研究では（1）情報を意味的に整理すること，(2）情報を時間的に整理すること，(3）ユーザの検索意図を明確にすることに焦点をあて，上記 3 つの要素を同時に考慮した Twitter を用いた災害時の検索システムを提案する．

本システムは，最初に文書要約手法 Maximal Marginal Relevance（MMR） [1] を利用して類似するツイートをグループ化し，情報を意味的に整理する．次に，各グループを話題が盛り上がった時期から新しい順に提示することで，情報を時間的に整理する．さらに，検索された複数のツイートの中からユーザが自身の検索意図に合うツイートを選び，選択した 1 つのツイートを新たなクエリとして再検索することで，対話的にユーザの検索意図と合致する情報を取得する．最後に，東日本大震災当時のツイートを利用して本システムのユーザ実験を行うことで，災害時のマイクロブログ検索における本手法の有効性を確かめる．

本稿では，2 章で関連研究を紹介し，3 章でツイートのグループ化と，ユーザのツイート選択による再検索手法について述べる．さらに 4 章で実験計画について説明し，5 章でまとめと今後の課題について述べる．

[1] www.biglobe.co.jp/pressroom/release/2011/04/27-1
[2] www.yomiuri.co.jp/net/security/goshinjyutsu/20110325-OYT8T00642.htm

2　関連研究

検索結果の意味情報の可視化は，検索結果に対するユーザの理解を支援する有力な方法である．Iwata ら [2] は，検索結果の複数の意味的側面を検索結果と共に可視化することで，検索結果を意味的に整理することの有効性を示した．White ら [5] は，Web 検索において検索結果一覧からユーザが閲覧したページの履歴を 5 つの観点で評価し，検索結果一覧よりもページ履歴の提示が検索結果の理解に役立つことを示した．O'Connor ら [4] はツイートを各話題毎にグループ化しながら Twitter を検索するシステムを提案している．彼らの手法は出現頻度の高いフレーズを含むか否かで検索結果のツイートをグループに分けている．一方，本提案手法ではツイートのテキスト全体を比較することでグループを作成する．高村ら [7] はツイートの要約を施設配置問題と捉え，文書の内容と時間的特性に注目して要約を行った．彼らは，文書の内容だけを用いる要約では，時間的に離れた異なるイベントについて言及しているツイートを混同してしまう可能性を指摘している．しかし，本論文ではグループ化の効果を調査することを目的としているため，時間的な特性については考慮しない．マイクロブログ検索については，TREC 2012 において Miyanishi ら [3] が，ユーザの選択したツイートをクエリとすることで検索結果を著しく向上できることを報告している．

3　マイクロブログ検索システム

本章では，Twitter からユーザの検索意図に適合する文書を検索するために，検索結果のグループ化と検索結果から選択したツイートを新たなクエリとして検索する手法を組み合わせる枠組について紹介する．

図 1: 提案システムのスクリーンショット

3.1 インタフェース

本研究で提案する震災時のマイクロブログ検索システムのインタフェースを紹介する．図 1 は提案システムのスクリーンショットである．まず，ユーザは画面上部の検索ボックス（図 1 の①）に関心のある話題についてのキーワードを入力して検索する．その結果，画面左に検索結果を表す列が出現し，キーワードに関係のあるツイートが，話題ごとにグループ化された状態で表示される（図 1 の②）．グループ内に表示されているツイートの数は，グループに属するツイートの総数を N とすると，$\log_3 N$ 件とした．よって，グループに属するツイートの総数に応じて表示するツイート数は対数的に増加する．これにより，表示させるツイート数を見ることで，話題の規模を大まかに知ることができる．また，表示中のツイートを右クリックするとグループのボックスが縦に広がり，そのグループに属するすべてのツイートを閲覧することができる．

さらに，検索意図に近いツイートをダブルクリックすると，そのツイートのある列の右に新たな列に再検索を行った結果が追加される（図 1 の③）．画面には検索結果が 3 列まで表示でき，3 列を超える検索結果は画面の左側に押し出される，しかし，押し出された列は消滅せず，図 1 の④の「prev」「next」をクリックすると列が右側や左側にスライドして過去の検索履歴を見ることができる．よって，ユーザは最後の検索結果からさらに再検索を行ったり，以前の検索結果から検索をやり直すことができる．

3.2 ツイートのグループ化

Twitter では，ある話題が盛り上がった時間帯に，その話題に関する用語を含むツイートが増加するという特徴がある [7]．よって，テキストの類似したツイートを 1 つのグループにまとめることでツイートを話題ごとに分けることができる．さらに各グループに含まれるツイートの量から話題の規模をおおまかに知ることができる．提案手法では，MMR を用いて全文検索結果のツイート集合からグループを代表するツイートを選択する．MMR は文書選択による要約手法であり，文書をクエリとの関連度と既に選択された文書との相違度の 2 要因でスコア付けして，スコアが最大の文書を選択する．ツイートを文書とみなすと，スコアは式 (1) によって与えられる．

$$\lambda Sim_1(d_i, q) - (1-\lambda) \max_{d_j \in S} Sim_2(d_i, d_j) \quad (1)$$

ここで，候補の文 d_i とクエリ q の関連度 Sim_1 には検索エンジンが出力した検索スコアを正規化して用いる．また，S は既に選ばれた文 d_j の集合であり，d_i と d_j の類似度 Sim_2 には文に含まれる形態素の idf を用いたコサイン類似度を使用する．λ（$0 \leq \lambda \leq 1$）はクエリと文書の適合を優先するためのパラメータであり，λ が大きいほど各グループ間の多様性が小さくなる．また，MMR では，指定した回数だけ式 (1) を用いて文書を選択することができ，選択した文書の数が作成するグループの数となる．グループの数はユーザが任意に指定できる．次に，選択されなかった文書を対象に，グループを代表する文書との Sim_2 を求め，それぞれ Sim_2 が最大となるグループに振り分ける．

3.3 ツイートをクエリとする再検索

本システムでは，検索画面に表示されているツイートをユーザがダブルクリックすると，そのツイートから抽出したクエリ語と元のクエリ語を組み合わせて再検索を行うことができる．再検索時のクエリの生成では，まず選択されたツイートを形態素解析し，固有名詞，一般名詞，およびサ変接続名詞のみを抽出する．さらに英数字のみで構成される 2 文字以下の語を除去し，残った形態素のうち，idf の上位 M 件をクエリ語集合 Q_t とする．検索ボックスに入力された元のクエリ語集合を Q_o とすると，再検索では，Q_t と Q_o それぞれから少なくとも 1 つのクエリ語を含むツイートだけを検索する．このように，元のクエリを選択されたツイートによって拡張することで，ユーザが新たに検

図 2: 実験に使用するシステムのスクリーンショット．A：フラットリスト形式（通常の検索方法），B：検索結果をグループ化した形式，C：グループ化＋マイクロブログ文書をクエリとする再検索（提案手法）．

索クエリを入力しなくても，よりユーザの意図に沿った文書を検索できるようになる．

4 実験計画

災害時の Twitter 検索における提案手法の有効性を評価するために，東日本大震災時の実際のツイートデータを用いた実験を行う．本実験では，下記に示す．グループ化の効果，再検索の効果，組み合わせの効果の 3 つの点について調べる．

- **グループ化の効果**：ツイートをグループ化することで，従来の方法に比べ，検索意図に合うツイートを効率良く収集できるか
- **再検索の効果**：選択したツイートをクエリとして再検索することで，元のクエリよりも検索意図に合致するツイートを発見することができるか
- **組み合わせの効果**：グループ化と再検索の組み合わせることでどのような効果があるか

4.1 データセット

本実験には，2012 年に開催された東日本大震災ビッグデータワークショップ[3]において Twitter Japan 株式会社から提供されたツイートデータを用いる．提供されたデータは，東北地方太平洋沖地震発生から 1 週間（2011 年 3 月 11 日午前 9 時 00 分〜同年 3 月 18 日午前 8 時 59 分）に発信された約 1 億 8000 万件の日本語ツイートである．それぞれのツイートは，ツイート ID，ユーザ ID，投稿日時，ツイート本文の 4 つの欄で構成される．

4.2 全文検索

提案システムでは，全文検索エンジン Lucene[4]を利用してツイートの索引付けと検索を行う．検索モデルには，ディリクレ平滑化を適用した言語モデル [6] を利用し，各文書に対するクエリ尤度を検索スコアとした．ここで平滑化パラメータ μ は経験的に $\mu = 2500$ とした．また，形態素解析には kuromoji[5] を使用した．

4.3 比較インタフェース

本実験では，図 2 に示す A,B,C の 3 つのインタフェースを比較する．図中の A は Twitter 社の公式検索インタフェースに代表される一般的な検索インタフェースと同様のフラットリスト形式で，検索結果のツイートをクエリとの適合度が高い順に羅列して表示する．適合度には検索エンジンが算出したスコアを用いる．図中の B は検索結果のツイートをグループ化して表示するだけのインタフェースである．A と B を比較することで，グループ化の効果を評価する．図中の C は提案システムであり，検索結果のツイートをグループ化して表示し，さらに検索結果のツイートを選択すると，選んだツイートを新たなクエリとして再検索することができる（図 2 の C）．B と C を比較することで再検索の効果を，A と C を比較することで組み合わせの効果を評価する．

[3]https://sites.google.com/site/prj311/

[4]http://lucene.apache.org/core/

[5]http://www.atilika.org/

表 1: 検索課題の例

初期クエリ	検索意図
大崎市　給水	宮城県大崎市で給水を行なっている場所が知りたい
石巻　安否	岩手県石巻市の安否情報が知りたい

4.4 検索課題

災害時の検索システムの性能評価を行うために使用する検索課題は，災害当時の情報要求を反映する必要がある．そこで，震災時の情報要求を表す質問ツイートに着目した．ここで，質問ツイートを,「おしえてください」「おしえて下さい」「教えてください」「教えて下さい」のいずれかのフレーズを含む何らかの情報を求めるツイートとする．質問ツイートの投稿者は Twitter を通じて質問に関する情報が得られると期待しているので，質問ツイートを Twitter 上での情報要求と見なすことができる．また，Twitter には他のユーザーのツイートをそのまま引用し，自身のフォロアーに投稿するリツイートという機能がある．リツイートされた回数はユーザ間での関心の高さを示すものと仮定できる．そこで，抽出した全質問ツイート 114,315 件の中から，1 日ごとに { その日にリツイートされた回数 }/{ 全期間にリツイートされた回数 } の降順で順位付けを行い，上位から順に

- 内容が震災に関係のある質問をしている
- 同じ内容の質問をしているツイートの重複は除去
- 提供データから答えが見つかることが確認できた質問を選択

の基準に従い，より多くの人にリツイートされた質問ツイートを選んだ．ここで，提供されたデータだけではツイート同士のリツイート関係を抽出できないこと，本実験では厳密なリツイート回数が必要ではないことから，文頭に「RT」と付くツイートをリツイートされたツイートとした．最後に選択した質問ツイートの内容を整形して検索意図とし，検索意図に適合する文書を見つけるための初期クエリを質問ツイートをもとに手動で作成した．最終的な検索課題は初期クエリと検索意図を 1 組とする．検索課題の例を表 1 に示す．

4.5 実験手順

ユーザ実験を行うため，本実験では 18 人情報科学の知識を持つ被験者を集め，1 人あたり 3 システムで各 2 タスク，合計 6 タスクの検索を行う．使用するシステムおよびタスクの順番は，ラテン方格を用いて公平に割り当てる．被験者には，タスク毎に初期クエリから検索し，その結果から検索意図に一致するツイートを 5 分以内に 5 件見つけるように指示し，制限時間内であっても検索課題に適合するツイートを見つけしだい終了とする．

4.6 評価指標

評価は，検索に要した時間，見つけたツイートと検索意図の関連度，被験者アンケートの 3 点で行う．時間は，検索開始から終了までの時間を計測して比較する．関連度は，3 名の評価者によって人手で判定された関連度を用いて比較を行う．アンケートは，各システムの使用後と実験全体の使用後に実施し，被験者の実際の感想をもとに評価する．

5 おわりに

本研究では，マイクロブログの検索結果を MMR を用いてグループ化し，さらに検索意図にあったマイクロブログ文書から再検索を行うシステムを提案した．さらに，東日本大震災時の Twitter のデータを使用した実験を行なっている．実験結果と考察は発表時に口頭で説明する予定である．

参考文献

[1] Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Rroducing Summaries. In *SIGIR*, pp. 335–336, 1998.

[2] Mayu Iwata, Sakai Tetsuya, Takehiro Yamamoto, Yu Chen, Yi Liu, Ji-Rong Wen, and Shojiro Nishio. AspecTiles: Tile-based Visualization of Diversified Web Search Results. In *SIGIR*, pp. 85–94, 2012.

[3] Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. TREC 2012 Microblog Track Experiments at Kobe University. In *TREC*, 2012.

[4] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory Search and Topic Summarization for Twitter. In *ICWSM*, pp. 2–3, 2010.

[5] Ryen W. White and Jeff Huang. Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs. In *SIGIR*, pp. 587–594, 2010.

[6] Chengxiang Zhai and John Lafferty. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *SIGIR*, pp. 334–342, 2001.

[7] 高村大也, 横野光, 奥村学. Summarizing microblog stream. 人工知能学会第 22 回 SWO 研究会 SIG-SWO-A1001-03, 2010.