1F4-6

ニュース記事の内容からの主題展開に関する研究

Subject Expansion from Content of News Article

任毅[*1] Yi Ren
佐藤真[*2] Makoto Sato
赤石美奈[*2] Mina Akaishi

[*1] 法政大学情報科学研究科 Graduate School of Computer and Information Sciences, Hosei University
[*2] 法政大学情報科学部 Faculty of Computer and Information Sciences, Hosei University

In this paper, we propose a method to expand the content about the subject of a news article. The method, first cluster to the collection of news articles by calculating the similarity between two articles. Then, it abstracts the keywords that can represent the main content of the clusters. At first, we collected experiment data, analyzed the text and classified the articles into groups. Then, we abstract the keyword and present the article groups in a timeline with the help of time stamp. We evaluated our method by experiment using 100 news articles and verified the effectiveness of the method.

1. はじめに

近年のコンピュータやインターネットの普及，更にスマートフォンアプリの発展に伴い，最新のニュース情報はニュースサイトに掲載されるだけではなく，スマートフォンアプリのプッシュ機能を通じて記事のリアルタイム性が極めて高くなっている．しかし，毎日膨大な量のニュース情報量が発信されるため，ニュースへ強い関心を持っていない読者にとって，ある記事を読んだ後，その事件の一つの側面を知ることができるが，その事件の全体像を把握することが難しい．あるサイトに関連記事が纏めて載せられている場合もあるが，事件の全貌を知るためには不十分である．検索エンジンではユーザは多数のタイトルなどを概観して情報を取捨選択する必要があり，文書数が増加した場合に大きな閲覧コストを要する．

そこで，このような問題を解決するために，ユーザが興味を持ったテーマに関する関連情報を提示する手法が認められる．本研究は，主に収集した話題に関する記事をトピックに基づき分類し，各トピックの出現の流れを提示する方法を提案する．これにより，読者の興味のある事件に対する関連記事のトピック展開を示すことができると考える．

言語処理分野において，大量文書の話題（キーワード）抽出する方法、又は文書の分類方法は多数提案されている．しかし単一話題の関連記事文書の話題整理方法は未だ提案されていない．

以下，第 2 章で本研究の関連技術について，第 3 章で本研究の提案手法の詳細を，第 4 章で本提案手法の有効性を調べるために行った実験とその結果を述べる．第 5 章で実験結果を考察し，最後に第 6 章でまとめを述べる．

2. 関連技術

大量文書からのキーワード抽出方法と文書分類の手法は多数提案されている．

[佐藤 05]は時系列ニュース記事における最新話題抽出方法を提案している．この方法では，記事間類似度の算出と記事クラスタリングを通じて話題を整理し，そして記事新鮮度と記事間類似度に基づく記事話題度を算出し，最新話題を抽出する．

[松尾 02]は，語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズムについて述べている．

[橋本 08]は，多数のトピックを含む文書集合に対して階層的クラスタリングを施し，クラスタ間の語彙使用の類似性に基づく構造化を行い，個々のクラスタについてこれを要約するキーワードおよび関係する主体を自動抽出することを提案している．

連絡先：任毅，法政大学・情報科学研究科，philia.china@gmail.com

3. 提案手法

本研究は，検索エンジンで取られたニュース検索結果を対象に話題整理する手法を提案している．

ある記事に関する情報が検索エンジンで簡単に手に入れられる．しかしながら，膨大な検索結果から，読者が記事の全体像を理解するのは，非常に困難である．

また，ニュースサイトは，最新の情報が掲載されているため，ひとつのニュース記事から，その事件の背景や原因、関連情報などを辿り，全体像を把握することは多大な労力を有する．例えば，4 月 6 日の記事「ボーイング 787、運行再開に向けた試験飛行完了」という記事には，787 型機の試験飛行のことしか書いてないため，その数ヶ月前の 787 型機の運航中止の事件には言及していない．そして，初めて 787 型機に関するニュースを読んだ読者にとっては，その主題に関して，どのような事件背景があるのかを理解することは難しいという問題がある．これを解決するための手法の提案が，本研究における目的である．

本研究の目的は，読者が興味を抱いた記事の事件に関する様々な情報をユーザに提示し，事件の全体像把握を支援することである．まず，ユーザが興味を抱いた事件に関するニュース記事を，検索エンジンにより収集し，各記事の日付、タイトル、本文を抽出し，一回の検索結果から得られた記事を一つのセットにする．

その後，記事群を提案手法でクラスタリングし，各々のクラスタの主題として，各クラスタの頻出キーワードを抽出する．

各クラスタに含まれる記事の日付に基づき，話題の展開と時間との関係を示す．

その後，各クラスタを代表すべき単語を抽出する．最後は分類結果を適当に簡素化すること，結果テキストが事件の全体を展示できるように検討している．

以下，研究のプロセスについて詳細に述べる．

3.1 記事文書集合の取得

分析対象とした文書は，主にウェブ検索エンジンで取得する検索結果の前の数十件の記事とした．主に記事の日付、タイトルと本文を抽出した．

3.2 前処理

本提案手法の前処理は，(形態素解析器)Chasen で形態素解析を行い，隣接の同じ品詞種類を持つ単語を組み合わせ，名詞、動詞(「動詞－非自立」除く)、記号-アルファベットの品詞種類を持つ単語を選び出す．また，記号やアルファベットなどの全角文字は半角文字に変換す．

3.3 記事のベクトル表現

記事集合のクラスタリングを行うために，記事間の類似度を算出する．類似する内容の記事は出現する単語の分布が類似しやすいと考えられる．しかし，同時に，事件を示す特定単語の頻度が非常に高くなる．例えば，現在の分析対象とした「787」(アメリカ合衆国のボーイング社が開発・製造する次世代中型ジェット旅客機)に関する記事では，「787」や、「ボーイングや」、「米」などの単語の頻度は非常に高いと示されている．そのため，この単語は記事間の類似度の計算とクラスタリングの結果に影響を与える．そこで，全記事数の半分以上を出現する単語と，一回だけ出現する単語を除き，類似度を算出した．

本提案手法の類似度の算出はベクトル空間モデルで文書をベクトル表現する，その際，代表的な単語の重み付け法の TF-IDF 法を用いる．そして本手法の記事数 N の記事集合中の記事 d_i の記事ベクトルを以下のように定義する．

$$d_j = (x_{j1}, x_{j,2}, \cdots, x_{iV}) \ (i = 1,2, \cdots, N) \quad (1)$$

$$tf_{i,j} = n_{i,j} / \sum_k n_{k,j} \quad (2)$$

$$idf_i = \log N / \{d: d \ni t_i\} \quad (3)$$

$$d_j = tf_{i,j} \times idf_i \quad (4)$$

式(2)の，$n_{i,j}$ は単語 i の記事 j での出現回数．式(3)の $\{d: d \ni t_i\}$は単語 i が出現する記事数を表す．

本定義を用い，任意の記事 i, j 間類似度 $S_{i,j}$を次式で求める．

$$S_{i,j} = \sum_{k=0}^{V} x_{i,k} \times x_{j,k} \quad (5)$$

3.4 記事クラスタリング

K 平均法を用いて，クラスタリングを行う．クラスタ数 n は経験的に$\sqrt[4]{N}$を設定する．クラスタに含まれる記事数が$\sqrt[4]{N}$以下になるまで，再帰的にクラスタ分割を行う．クラスタリング手順を以下に示す．

ランダムに一つずつの記事を n 個のクラスタの中心とするクラスタを生成する．

(1) それぞれのクラスタの中心を計算し，中心ベクトルを生成する．

(2) 他の記事と各ベクトルとの類似度を計算し，最大類似度のクラスタに割り当てる．

(3) 割り当てが終了した後，各クラスタの中心ベクトルとの類似度が最も高い記事を各クラスタの中心として(2)以降を繰り返す．但し，もし一つの記事しか含まないクラスタがあれば，不平衡とする，最大のクラスタの中心との距離が遠い記事を切り離し，このクラスタの中心とする．

(4) もしクラスタリングの結果が前回と変わらなければ終了となる．もしクラスタの記事数が$\sqrt{N}$より大きければ，このクラスタを一つの記事集合として更にクラスタリングを行い，(1)以降を繰り返す．

3.5 キーワード抽出

本提案手法のキーワードは，各クラスタの内容を代表する単語である．

各クラスタのいくつかの記事で，単語 i の出現頻度は以下の通り定義する．

$$a_{i,j} = \begin{cases} 1 & (\text{単語 i は文書 j に出現する}) \\ 0 & (\text{単語 i は文書 j に出現しない}) \end{cases} \quad (6)$$

$$f_i = \sum_j a_{i,j} \quad (7)$$

各クラスタの最上位単語 3 つを抽出し，キーワードとする．

3.6 時間軸

各クラスタに含まれる記事の日付情報から，各クラスタ内に含まれる記事の最初の日付と最後の日付を取り出し，クラスタの出現時間帯とする．これを基に，クラスタの記事の出現時間帯を時間軸に沿って表示する．

これにより，各クラスタの記事が示している時間帯が視覚化される．

4. 適用例

4.1 実験条件

提案手法の有効性を確認するために，2013 年 1 月 25 日にグーグル(Google)から収集した 100 件の記事を対象として本手法を適用した．

4.2 クラスタリング結果とキーワード抽出結果

表 1 の記事のクラスタリング結果を木構造で示す．左側の列は各クラスタに対応している木ノードの深さ，そしてノード間の親子関係を示している．

表 1 の中央部には，記事のタイトルを示し，右側の列は各クラスタの頻出キーワードを示した．

4.3 時間軸表示結果

図 2 に，クラスタの出現期間を時間軸に沿って示す．図 2 によると，1 月 25 日に収集した記事集合の日付は 1 月 11 日から 1 月 28 日までとなる．

5. 考察

分析対象データをクラスタリングした結果，記事集合は 19 個のクラスタに分割された．例えば，$C_{1,1,1}$ では，キーワードとして「見通し」、「全日本空輸」、「設計」を抽出した．このことから 787 型機の未来に関すること、ANA に関する情報、787 の設計問題、この 3 つのことを書いた記事のグループであると推測できる．

しかし，小さなクラスタの場合，キーワードだけで内容を推測することは，かなり困難といえる．例えば，$C_{1,3,2}$ には，2 つの記事しか含まれていない．このため，単語の頻度だけでは，記事グループの内容を正しく示すことは難しいといえる．

更に，グループのキーワード抽出結果に，日付の単語がいくつか現れる．このことから同じ日にその事件の関連事件が発生したことがわかる．例えば，$C_{1,3,1}$ の 6 つの記事からのキーワードには，「7 日」と「27 日」があり，1 月 7 日と 1 月 27 日に 787 に関する事件が発生したことを推測できる．そして記事内容を読むと，7 日に発火事件，27 日に国際便の欠航が多数発生したこと分かる．

図 2 の時間軸の表示結果から見ると，記事収集時点と近い記事の数がより多いことがわかる．そして各記事グループが占める時間帯は主要内容によって違うことがわかる，例えば，ANA、設計、見通しを説明する記事グループ $C_{1,1,1}$ が 1 月 11 日から収集時点まである，つまり，このグループ情報は一時的なことではないことがわかる．更に，例えば，$C_{2,1}$ と $C_{2,2}$ は原因の

1	2	3	4	記事記号, 記事タイトル		キーワード
C_1	**$C_{1,1}$**	$C_{1,1,1}$		92 : 20130124 アングル:B787運航停止、航空大手2社への業績影響は限定的 0 : 20130125 焦点:緊急着陸したANAの787型機、原因究明さらに難航も 2 : 20130129 ボーイング787、トラブルの原因解明は長期化の様相!?「夢の飛行機」に待ち構えていた思わぬ落とし穴 3 : 20130128 ボーイング、787の運航停止コストは最大50億ドルとの見方も 38 : 20130111 ボーイング787の包括調査へ 米航空当局 53 : 20130124 キャンセルの可能性低い=787で-米アナリスト 76 : 20130121 米運輸安全委、787型機めぐり充電装置メーカーなど調査へ 79 : 20130123 ANA国内線サマーダイヤを決定、成田~広島などを新設、787使用再開時期によって変更も 93 : 20130117 欧州航空安全庁、「787」運航停止の米当局命令受け入れ		見通し=4.0 全日本空輸=3.0 設計=3.0
		$C_{1,1,2}$		14 : 20130123 原因は“ボーイング社による”手作業ミス? 787機トラブルで日本企業は濡れ衣を 21 : 20130117 ボーイング787、世界で運航停止へ 日米当局が命令 70 : 20130115 出火や燃料漏れ…787トラブル続々、今月6件 71 : 20130112 米ボーイング787型機の安全性問題、外注などが原因でない=民間航空機部門責任者 86 : 20130120 新たな設計、危うさ潜む ボーイング787 軽量化の徹底、大幅な電気化 91 : 20130117 米連邦航空局、「Boeing 787」機の運行停止を命じる 95 : 20130123 787 新千歳でも欠航続く		安全性=4.0 全日本空輸=3.0 国土=3.0
		$C_{1,1,3}$		83 : 0 エアバス CEO:「A350」はボーイング 787 よりリスク低い 37 : 20130122 ボーイング 787、早期解決は困難か 米国で再検査へ		説明=2.0 米国=1.0 22日=1.0
	$C_{1,2}$	$C_{1,2,1}$		C1,2,1,1	22 : 20130123〔焦点〕米ボーイング 787 型機電池トラブル、FAA の認可の適切性めぐり議論 29 : 20130125 ボーイング 787 運行再開メド立たず!トラブル原因不明で長引く飛行停止・出荷停止 34 : 20130125 ボーイング 787 運行停止で 3 日間で計 88 便欠航へ…ANA 44 : 20130121 ANA、787 型機の運航見合わせを 28 日まで延長 54 : 20130123 東レ、787 の収益浮揚力に乱気流 88 : 20130110 ANA の成田―サンノゼ線、計画通りボーイング 787 で就航へ	高松=4.0 航空会社=3.0 全日本空輸=3.0
				C1,2,1,2	87 : 20130110 ANA機でも787の燃料漏れ起きていた すでに改修、JAL機と合わせて8機で不具合発見 31 : 20130118 運航停止の787、問題解決と再開のメドは?	航空会社=2.0 必要=2.0 指示=2.0
				C1,2,1,3	10 : 20130128 ボーイング 787 運行停止で ANA&JAL が大ピンチに? 96 : 20130121 全日空はボーイング 787 の運行停止を延長と伝わるが底堅い	現在=2.0 7 日=2.0 運行停止=2.0
		$C_{1,2,2}$	$C_{1,2,2,1}$	4 : 20130129「787」電池により厳格な審査求めていた業界団体 16 : 20130126 日米の787トラブル 電池の製造時期にずれ 18 : 20130124 バッテリー分解調査に着手 787発煙で安全委 20 : 20130123 ボーイング、新型機「787」で窮地 24 : 20130126 電気系統「専門委員」の参加検討 787トラブル調査 58 : 20130121 GSユアサに立ち入り検査 全日空787トラブルで 61 : 20130121 全日空、23~27日に141便欠航「787」停止長期化も 65 : 20130124 バッテリー問題の原因、まだ不明=787型機トラブルで-米航空局長官 75 : 20130122「787」の夢と現実		23 日=4.0 安全性=3.0 設計=3.0
			$C_{1,2,2,2}$	60 : 20130124 革新にリスクはつきもの「787」で苦悩するボーイング 11 : 20130126 全日空、2月は379便が欠航 787運航停止で 36 : 20130118 運航停止が脅かすボーイング787の夢 82 : 20130122 日本がボーイング787を大量購入した理由 84 : 20130117 米ボーイング技術者労組がスト回避へ、787型機への対応優先		同社=4.0 見通し=3.0 ドリームライナー=3.0
	$C_{1,3}$	$C_{1,3,1}$		35 : 20130123 リチウム電池の評価なぜ高いのか-787型機バッテリーに関する科学 9 : 0 米ボストンの787バッテリー発火事故、原因は依然不明=NTSB 15 : 20130125 787のバッテリーで熱暴走、原因は依然不明 米運輸安全委 32 : 20130126 全日空、2月1日以降379便欠航 787トラブル 43 : 20130116 ボーイング787不具合相次ぐ 出火バッテリーの写真公開 62 : 20130121 全日空、23~27 日に 141 便欠航「787」停止長期化も		7 日=3.0 27 日=3.0 今月=2.0
		$C_{1,3,2}$		19 : 20130127 全日空ボーイング787運航停止 計379便が欠航 81 : 20130122 ボーイング787型機		全日本空輸=1.0 高松=1.0 見通し=1.0
		$C_{1,3,3}$		77 : 20130110〔アングル〕米ボーイング787型機のトラブル、リチウムイオン電池の安全性への議論高まる 52 : 20130108 米当局がボーイング調査へ 787発火事故で 56 : 20130124 ボーイング787、運航再開めど立たず 59 : 20130123 運行停止のボーイング787、トラブル集中の理由とは 94 : 20130112 ボーイング CEO「調査で安全性裏付ける」787トラブルで声明		今回=3.0 7 日=3.0 2011 年=3.0
…	…	…		…		…

表 1 2013 年 1 月 25 日に収集した 100 件 787 の検索結果によりクラスタ樹(部分)及び要約キーワード上位 3 語

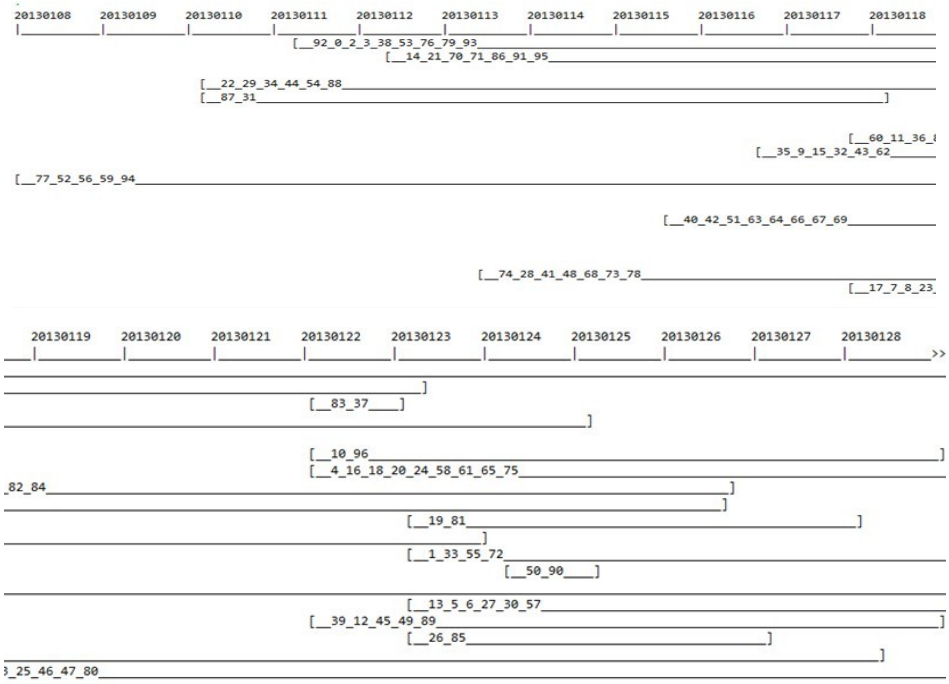

図 2 クラスタ樹の各クラスタが時間軸に表示する結果

具体的な調査に関する記事であり，時間線は時間軸が示しているように $C_{1,3,1}$ より短いことがわかる．

また，今回の実験結果の問題点としては，以下の 2 点があげられる．

1 つは，小さなクラスタの効果は悪くなるということ，つまり，どんな側面を反映するのかがが明らかになっていないということである．

2 つ目はいくつかのキーワードが複数のクラスタに存在することである．但し，この現象は正常だといえる，しかし，複数のクラスタの要約キーワードの重なるところが複数ある場合，クラスタの特徴として，あるの記事には記事の複数の側面に関する情報が載っている可能性が高いといえることである．

例えば，$C_{2,3}$ と $C_{3,1,1}$ の最上位の要約キーワードは全て同じだという現象がある場合，クラスタリング手法を再び検討する必要がある．

6. まとめ

本論文では，ニュース記事の主題に関する情報展開の手法を提案した．まず前処理と記事ベクトルの計算について述べた．次に本論文の核心である 1 つの主題に関する記事集合のクラスタリング手法と結果の要約キーワード抽出手法を説明した．実際にこの手法を 787 事件に適用し，予想通りの結果を得ることで，記事集合の事件に関する情報を分類する事例を示した．

本稿の提案手法の利点は，記事集合から事件に関する情報を分類することだけではなく，各種類の要約キーワードを抽出ことにある．そして各種類を時間軸に表示することから事件に関する情報の発生前後を知ることが出来る．

しかしながら，クラスタリングの結果として各クラスタの要約キーワードが重なる現象があった，そして小さなクラスタの結果がかなり悪いことが現在の問題点となる．

今後の課題としては，現時点で読者に向けた評価実験を行わなっていないため，本提案手法の実の効果が未だ明らかになっていないためない．アンケートの形式で読者を対象に評価実験を行うことが挙げられる．それに，検索エンジンがリアルタイムの特徴を持つため，単にある時点で収集する記事集合を対象とするだけでなく，時間の変化に伴い収集する記事集合から，主題事件の発展などの情報を明確化することが挙げられ，それについての検討が必要である．

参考文献

[佐藤 05] 佐藤吉秀，川島晴美，佐々木努，奥雅博: “時系列ニュース記事における最新話題抽出方法”，電子情報通信学会，信学技報 NLC2005-1 ,2005.

[湯浅 95] 湯浅夏樹，上田徹，外川文雄: “大量文書データ中の単語間共起を利用した文書分類”，情報処理学会論文誌，Vol.36 No.8 , 1995.

[橋本 08] 橋本泰一，村上浩司，乾孝司，内海和夫，石川正道: “文書クラスタリングによるトピック抽出および課題発見”，社会技術研究論文集，Vol.5，216-226 , 2008.

[松尾 02] 松尾豊，石塚満，: “語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム”，人工知能学会論文誌，Vol.17，3D，2002.

[深谷 5] 深谷亮，山村毅，工藤博章，松本哲也，竹内義則，大西昇: “単語の頻度統計を用いた文章の類似性の定量化”，電子情報通信学会論文誌，Vol.J87-D-II，No.2，pp.661-672，2004.

[佐藤 6] 佐藤吉秀，川島晴美，佐々木努，大久保雅且: “文書の類似度と新鮮度に基づく話題語抽出”，情報処理学会研究報告，2005-NL-165，2005.