様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　6 月　17 日現在

研究種目：若手研究（スタートアップ）
研究期間：2008 ～ 2009
課題番号：20860085
研究課題名（和文）文書集合における潜在的意味に着目した特徴量選択手法の提案について

研究課題名（英文）The feature selection focused on a latent semantic in a document set

研究代表者
横井　健（YOKOI TAKERU）
東京都立産業技術高等専門学校・ものづくり工学科・助教
研究者番号：40469573

研究成果の概要（和文）：本研究では、大規模な文書集合における潜在的意味（トピック）に着目し、そのトピックを用いて文書情報を表現する特徴量を選別する手法を提案した。特に、大規模な文書集合からトピックを取得するために、分割したそれぞれの文書集合から得られたトピックを結合することで、もとの大規模な文書集合を直接扱う場合に近いトピックを取得する手法を提案した。また、その手法によって得られた特徴量をユーザの興味情報による情報フィルタリングに応用することで、その特徴量の有用性を検証した。

研究成果の概要（英文）：The novel feature selection, i.e. an index words selection, based on a latent semantic (topic), for a large document set has been proposed in this study. The combination of topics among divided document sets was especially focused to obtain topics from a large document set, so that our proposal could extract a similar topic with the topic from a large document set with an extraction directly. In addition, the index words selected based on the above topics are confirmed to be efficient by the application of an information filtering with a user's interest.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,330,000	399,000	1,729,000
2009 年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	2,530,000	759,000	3,289,000

研究分野：工学
科研費の分科・細目：電気電子工学・システム工学
キーワード：テキストマイニング、トピック抽出、特徴量抽出、自然言語処理

1.　研究開始当初の背景

昨今の情報化の流れの中で、数多くの文書情報がインターネットに代表されるネットワークを通じて公開されており、誰でもそれらの文書情報にアクセスすることが可能となっている。しかしながら、公開されている情報量が膨大で、多くの一般ユーザにとって、必要な情報に辿り着くことが困難となってきており、それらの膨大な情報の中から必要な情報を抽出、または、整理する技術の開発が急務である。それらの技術では、それぞれの文書情報を特定する特徴量の抽出が必要

である。

従来、文書情報の特徴量を縮約、または、選別する手法としては、単語に対する重みを要素とするベクトル（ベクトル空間法）で文書を表現し、その単語の重みによる特徴空間を、別の特徴空間へ写像する手法が一般的である。文書を表現する特徴空間の変換を行う手法では、その特徴空間の基底をある文書集合に基づいて決定する。したがって、文書集合の規模が大きくなると、写像先の特徴空間の決定に膨大な計算が必要となる。また、文書を評価する際には、その特徴空間へ文書を写像して評価を行う必要がある。

一方、文書集合中から重要なキーワードを抽出する研究も行われており、その中のひとつとして、文書、または文書集合中に高頻度に出現する単語を基準とし、それらの高頻度語との関係性から、重要な語を判定するという手法が提案されている。この手法は、文書集合が大規模であっても比較的少ない計算量でその単語の重要度を計算できるため、WWW 等の大規模な文書集合にも適している。しかしながら、高頻度語でない単語に関しては重要ではないと判断されるため、文書集合中において関連する文書が少ないトピックに含まれる単語は選別から漏れてしまう可能性が含まれている。

2. 研究の目的

本研究では、1 節で述べた研究背景を踏まえ、次の 2 つの事項について検討する。

(1) 潜在的意味に基づいた特徴量抽出。
(2) 文書フィルタリングや文書分類の精度改善を目指した特徴量選別。

以下、上記 2 つの事項について詳しく述べる。

まず、(1)では、文書集合中に含まれる潜在的意味、特にここではトピックと呼ばれる概念に基づいて、単語そのものの重要性を評価し、その重要度に基づいて単語の選別、つまり文書を表現する特徴量の選別を提案する。トピックに基づいて単語の重要度を決定することで、大規模な文書集合において関連する文書の少ないトピックに含まれる重要語も抽出が可能になると考えられる。さらに、この手法では、文書を構成する特徴空間は単語に対する重みを基底とした空間で構成されているため、従来手法のように空間を変換することなく特徴量の選択や縮約を行うことが可能となる。特に、(1)を提案するにあたっては、大規模な文書集合からのトピック抽出方法の効率化が必要である。

次に(2)では、文書を特定するための適切な特徴量の選択を実施することで、文書フィルタリングや文書分類において精度改善が見込まれる。そこで、本研究では特に、ユーザの興味を用いた情報フィルタリングに選別された特徴量を適用し、その効果を検討する。

3. 研究の方法

本研究では 2 節で述べた 2 つの事項について検討を行う。まず(1)の「潜在的意味に基づいた特徴量抽出」では、次の 5 項目について提案と検討を行った。

(1) 文書集合からのトピック抽出
(2) トピックからそれらのトピックを表現するキーワードの選別
(3) キーワードをもとに単語の絞り込み
(4) 得られた単語の評価
(5) 大規模な文書集合への適用

まず、(1)では、従来、的確に文書集合からトピックを抽出できると報告されている文書-単語行列に対する NMF (Non negative Matrix Factorization)によるトピック抽出手法を用いた。

(2)、(3)では、高頻度語と他の単語の関連性による重要語抽出手法で用いられている単語間の関連性を、トピックを表現するキーワードとその他の単語間に適用し、トピックを基に重要な単語を選別する手法を提案した。NMF によって得られるトピックは文書集合中に含まれる単語の重みで構成されており、その重みの高い単語がトピックを表現する上で重要な単語だと判断した。また、そのトピックを表現する上で重要と考えられる単語だけでなく、その周辺に共起する単語も文書を表現する特徴量としては有用と考え、トピック中で重みの高い単語とその他の単語のカイ二乗値に基づいて最終的な特徴量（単語）の選別を実施した。

一方、 (4)においては、実際の文書集合に上記手法を適用し、得られた単語について検討を行った。

最後に(5)で、大規模な文書集合から NMF を用いてトピックを抽出するために、文書集合の分割とトピックの結合によるトピック抽出の効率化を提案した。また、提案手法によって得られるトピックについて大規模な文書集合に対して直接 NMF を適用してトピックを抽出する従来手法との比較検討を行うことで、効率化の有用性を検証した。

次に、研究目的事項(2)の「文書フィルタリングや文書分類の精度改善を目指した特徴量選別」についてであるが、次の 4 項目について実施、検討を行った。

(1) 研究目的事項(1)で述べた手法により選別された特徴量で文書を表現
(2) 再表現された文書集合を用いた文書フィルタリング・文書分類
(3) 単語の削減度合い等のパラメータについての検討
(4) 精度の検討と本研究で提案した特徴量選択手法の有効性についての検証

まず、(1)では、提案した特徴量選択手法に

よって選択された単語を用いて、新たな単語-文書行列を作成した。

(2)では、主にユーザの興味情報（ユーザプロファイル）に基づいた文書フィルタリングを実施した。ユーザプロファイルは文書をベクトル空間法を用いて構成した際に良く用いられている Rocchio の関連フィードバックに類似した学習文書の重み付き重心をユーザプロファイルとした。

(3)では、単語選別に用いるトピックの抽出数や基準となるトピックを表現する単語数、共起情報に基づいて追加する単語の数について焼き鈍し的に検討を行った。

最後に(4)では、本研究で選別された単語により構成された空間で表現した文書集合を用いた場合における(2)で述べたユーザの興味情報に基づいた文書フィルタリングの精度と、選別を実施しない場合、潜在的意味解析として一般的な LSA(Latent Semantic Analysis)によって特徴空間を縮減した場合の比較を行い、本特徴量選択手法の有用性について検証を行った。なお、比較はテストコレクションを用いて実施した。

4. 研究成果

まず、本研究で提案した特徴量選択手法によって得られた単語が構成する空間において実施した文書フィルタリングの結果について述べる。図 1 はテストコレクション MEDLINE における Query No.1 に該当する文書集合を用いてフィルタリングを実施し、精度を比較したグラフである。MEDLINE テストコレクションでは、1,033 の文書が用意されており、Query No.1 では、そのうち 37 件が関連有りとラベル付けされている。また、含まれる全単語数は 7,014 語である。

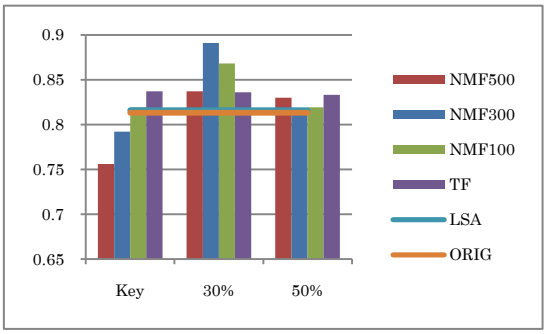

図 1 フィルタリング精度の比較

図 1 において、NMF500、NMF300、NMF100 はそれぞれ提案手法を表しており、NMFxxx の xxx が文書集合から抽出するトピック数を表している。TF は、トピックの重要単語ではなく文書集合中における高頻度語をもとに単語選別を行った場合の結果である。また、Key、30%、50%は、元々の文書集合に含まれる全単語数のうち選択する単語数の違いを示しており、Key はトピックから得られた単語のみを用いた場合、30%、50%は、Key で得られた単語に加え、カイ二乗値を用いて共起する単語も追加して得られた単語数の総単語数に占めるおおよその割合をそれぞれ示している。(例えば、30%であればトピックから得る単語とカイ二乗値によって得る単語数の総数を 7,014 の 30%、つまり 2,100 語程度に設定する。) さらに LSA、ORIG は LSA によって空間縮減を行った際のフィルタリング精度、前処理を行わないでフィルタリングを実施した際の精度をそれぞれ表している。この図より、トピックによって得られる重要語だけを用いた場合は、ORIG や従来手法の LSA に比べても精度の改善が見られない。したがって、トピックから取得する単語だけでなく、それらと共起する単語を取得することは有用であると考えられる。また、NMF300 の 30%が一番良いフィルタリング精度を実現しており、適切なトピック数と適切な選別単語数を設定することができれば本提案手法が有効であることが分かった。なお、このとき、5,000 語程度の利用特徴量の削減が実現できている。

次に、大規模文書集合からトピック抽出を効率化するために実施した提案手法の結果として、オリジナルの文書集合そのものから得たトピックと本研究で提案した分割文書集合からトピックを抽出し、その後そのトピック群を結合することによって得られたトピックの一致割合を図 2 に示す。図 2 では、1 週間各日の新聞記事集合(各日 50 文書から 100 文書程度) の 7 日間の文書集合において 2 日分の文書集合を組み合わせ、組み合わせた文書集合から直接抽出したトピックとそれぞれから抽出し、結合したトピックの一致割合の平均値を示している。

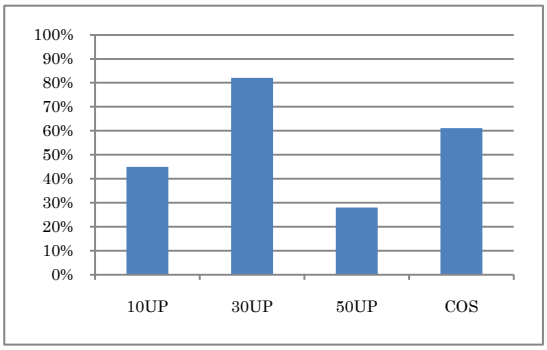

図 2 トピックの一致割合

図 2 において、10UP、30UP、50UP はそれぞれ、比較トピックにおける上位 20 単語中 10 単語以上が合致しているトピックの割合、上位 100 単語中 30 単語以上ないし 50 単語以上が合致しているトピックの割合をそれぞれ示している。また、COS はコサイン類似度が 0.7 以上であれば一致していると判断した

場合のトピックの一致割合を示している。10UP では、比較する単語数が少なすぎることが原因で、一致度合いが低くなっていると考えられる。一方、50UP では、50 単語以上という制約が厳しすぎたため、一致割合が下がったと考えられる。一方、30UP においては、80%以上のトピックが合致していると判断できた。これより、上位単語の 3 割程度は多くのトピックで同じ単語が出現していることが分かった。また、分割した文書集合からそれぞれトピックを抽出することによる処理時間の比較を行った。その結果、本提案手法により、大きな文書集合から直接トピックを抽出する場合に比べて、1/2 から 1/8 に処理時間を短縮することができた。また、より大規模な文書集合において本提案手法の有用性を検証するために、文書数約 4,700 件、総単語数約 42,000 語の cluto と呼ばれる文書分類評価のためのテストコレクションを用いて検証を行った。検証実験では、それらを 1,200 件程度の 4 つの文書集合に分割し、本手法の検証を実施した。その結果、直接大規模な文書集合から求めたトピック（従来手法）のうち 80%程度のトピックを提案手法によって抽出することができた。表 1 に上記 cluto の実験におけるそれぞれの手法で取得したトピックの代表語（トピック中の最も重みの大きい語）例とその語のトピック中における重みを示す。また、色付きの単語は提案手法と従来手法に共通して出現する代表語を示

表 1 取得トピックにおける代表語とその重みの比較

No.	従来手法		提案手法	
	単語	重み	単語	重み
1	library	0.852	title	0.778
2	system	0.509	service	0.751
3	inform	0.323	inform	0.722
4	science	0.119	cost	0.598
5	base	0.117	index	0.532
6	term	0.115	system	0.479
7	journal	0.108	catalog	0.477
8	catalog	0.108	journal	0.469
9	search	0.098	class	0.462
10	provide	0.081	search	0.449
11	service	0.076	research	0.441
12	literature	0.073	retrieve	0.428
13	cost	0.062	library	0.423
14	language	0.062	term	0.396
15	paper	0.058	data	0.383
16	research	0.039	revel	0.371

している。
表 1 に示した代表語からも提案手法と従来手法ではかなりの割合で同種のトピックが取得できていることが確認できた。

本研究では、文書集合に含まれるトピックに基づき特徴量を選別する手法を提案した。従来、特徴量の選別は人手による選別や頻度情報などの統計的情報に基づいた選別が主であった。また、特徴量の縮約という意味では、空間変換が主だった手法であった。一方、本研究で提案した特徴量選別手法は、トピックという文書個々が持つ単語よりも上位の概念に着目し、それに基づいて特徴量の選別を行うという点で、学術的にも新しい視点を得たと考えられる。また、本研究の成果は、昨今の WEB 上の文書群に代表されるような大規模な文書集合において、それらの文書集合から効率的にトピックを抽出する際に有効性を発揮すると考えている。さらに、それらのトピックを利用した重要語の抽出手法を提案したことで、大量の文書情報を整理、活用するための一助になると思われる。

5. 主な発表論文等

〔雑誌論文〕（計 1 件）

① T. YOKOI, H. YANAGIMOTO and S. OMATU: “Information Filtering using Index Word Selection based on the Topics”, International Journal of Information Technology, Vol. 5, No. 2, pp.81-87, 2009, 査読有.

〔学会発表〕（計 2 件）

① T. YOKOI and H. Yanagimoto: “Topic Extraction from Divided Document Sets”, The 5th International Conference on Web Information Systems and Technologies (WEBIST2009), 2009, Book of Abstract p.50, Proc. pp.661-666 (CD-ROM).

② T. YOKOI and H. YANAGIMOTO: “Topic Extraction for a Large Document Sets with Topic Integration”, The 3rd International Conference on Knowledge Discovery and Data Mining (WKDD2010), 2010, Proc. pp.46-49.

6. 研究組織

(1)研究代表者

横井　健（YOKOI　TAKERU）
東京都立産業技術高等専門学校・ものづくり工学科・助教
研究者番号：40469573