公開

（事後評価・研究者用）

研　究　報　告　書

「セマンティック情報を用いた情報検索システム」

研究期間：平成30年10月～平成31年3月
研究者番号：50138
研 究 者：欅　惇志

1．研究のねらい

本課題では，Web 検索を想定した，語彙情報やクエリのメタ情報を用いたランキング学習への適応を行った．ランキング学習とは，機械学習技術の一種である（半）教師あり学習を用いて，ユーザによって与えられた検索質問（クエリ）に対して正解となる文書を適合度の降順に並べる技術の総称である．既存のランキング学習では，過去に情報検索に関する研究分野にて培われてきた，各種スコアリング手法（単語の重み付け，文書の性質分析など）や各種情報（ユーザ属性，検索履歴など）を特徴量として用いて学習モデルを構築している．その際の課題として，課題（1）単語を記号して扱い，その意味を考慮していない，課題（2）すべてのクエリに対して単一の学習モデルを構築する，課題（3）クエリに対して学習データが存在しない状況を想定していない，が挙げられる．

課題（1）に関して，意味を考慮することで，ユーザが意図していない内容を含む文書を検索結果から除外したり，ユーザが潜在的に求めつつもクエリとして表現できていない内容を含む文書を検索結果として提示したりすることができるようになることが期待される．

課題（2）では，情報検索に関する過去の研究の知見として，検索タスク（人名検索や QA 質問，単語の定義検索やローカル検索など）ごとに異なる検索方針を設計することでより高精度な検索が実現できるということが報告されている．このことからも，ランキング学習においても検索タスクごとに学習モデルを構築することでより高精度な検索が実現できることが期待される．

課題（3）において，一般的に Web 検索では非常に多用なクエリが発行され，全体のうち 15% のクエリは 1 度しか発行されていないクエリだと報告されている．このような状況において，蓄積された学習データ中に同一クエリが存在することを仮定することは現実と乖離しているといえる．それに対して，既存のランキング学習では，未知クエリが発行される状況を想定されていないため，未知のクエリに対しても高精度な検索を目指すことはより実用的な技術である．

これらを踏まえ，本研究では上記の三つの課題に取り組むことで，より高精度かつ実用的なランキング学習の実現を目指した．

2．研究成果

（1）概要

提案手法について述べるうえで基本的な技術となるランキング学習について述べる．ランキング学習の概要を図 1 に示す．ランキング学習では，文書集合と，クエリ，クエリに対する

文書の適合性評価を持つデータセットを用いてモデル構築を行う．その際，文書（文書特徴量）やクエリ（クエリ特徴量），それらのペア（文書-クエリ特徴量）から各種統計量などを算出したものを列挙した特徴量ベクトルを作成し，クエリに対する文書の適合度を教師データとして付与する．これら特徴量ベクトルと教師データから学習モデルが構築された後は，通常の Web 検索と同様にユーザが入力したクエリに対して検索結果である順位付きリストを提示する．その際，文書集合は膨大な量となるため，すべての候補文書に対してスコアリングを行うことは非常に高コストとなる．従って，まずは単純なスコアリング手法を適用して得られた上位 k 件の文書（Top-k 文書）を抽出し，それらに対してリランキングを行うことが一般的であり，本課題でもその方針に倣う．

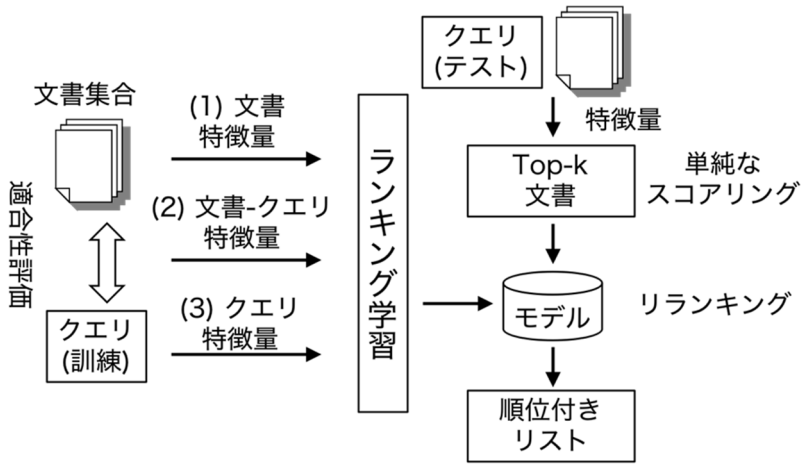

図 1 ランキング学習のフレームワーク

これらを踏まえ，本課題の成果として，研究テーマ (1) Web クエリに対する語義の曖昧性解消技術の提案，研究テーマ (2) タスクごとの適合情報の傾向の分析，研究テーマ (3) テストクエリと同一クエリが学習データに含まれない状況を想定したランキング学習手法の提案，を行った．

研究テーマ (1) 関して，課題 (1) で触れた単語の意味を考慮する際に，近年は，各単語を(単語数と比較して）低次元の固定長次元のベクトルで表現する分散表現を用いることが一般的である．文書やクエリをそれぞれ doc2vec と query2vec と呼ばれる分散表現で表現すると，意味的に近い文書･クエリが意味空間中で近接する特徴を持つ，単語と同一次元数のベクトルとして構成することができる．つまり，分散表現は，文書やクエリを大域的な観点における意味的距離を計測するアプローチである．その一方で，自然言語処理の分野では，単語単位における意味を考慮する際に語義の曖昧性解消と呼ばれる技術が多用されてきた．これは，自然言語の持つ曖昧性を解消する技術であり，例えば，jaguar という単語の主要な語義（語の意味）としては動物とカーブランドがあり，ある文書中の jaguar がいずれの意味であるのかを周辺の語の生起状況から特定する．特に Web クエリは，構成す

る単語数が高々数語程度であるため，一つずつの単語の持つ影響力は大きい．従って，本研究では，単語の大域的な意味を考慮するための分散表現と，単語の局所的な意味を考慮するための語義情報を用いる．その際，クエリに対する語義の曖昧性解消技術は存在しないため，本課題にて提案を行った．

研究テーマ (2) において，課題 (2) で述べた通り，検索タスクごとに学習モデルを構築することは高精度検索において有用であると考えられるが，検索タスクに依存せずに高精度検索に寄与する特徴が存在することが想定されるため，検索タスクごとに完全に独立した学習モデルを構築するのではなく，特徴量ベクトルの一つの特徴量として付与することでより柔軟なモデル構築を行った．

研究テーマ (3) では，課題 (3) に関して，モデル構築に用いた訓練データに存在しないクエリに対しても高精度な検索が実現できることを目指して，クエリは訓練クエリとテストクエリに分割した．従って，クエリから得られる特徴量だけから適合文書を知る手がかりは得られず，上手く文書-クエリ特徴量の設計を行うことで，適合文書の特定を目指す．

(2) 詳細

研究テーマ (1)

本課題では，文書とクエリの局所的な意味の類似度計測のため，それぞれの持つ単語の語義の比較を行う．なお，自然言語処理分野では，語義の曖昧性解消において，IMS と呼ばれるツールが多用される．これは文単位を入力として想定されており，例を図 2 に示す．入力文のうち曖昧性を持つ単語に対して語義の候補とその確率を付与する．

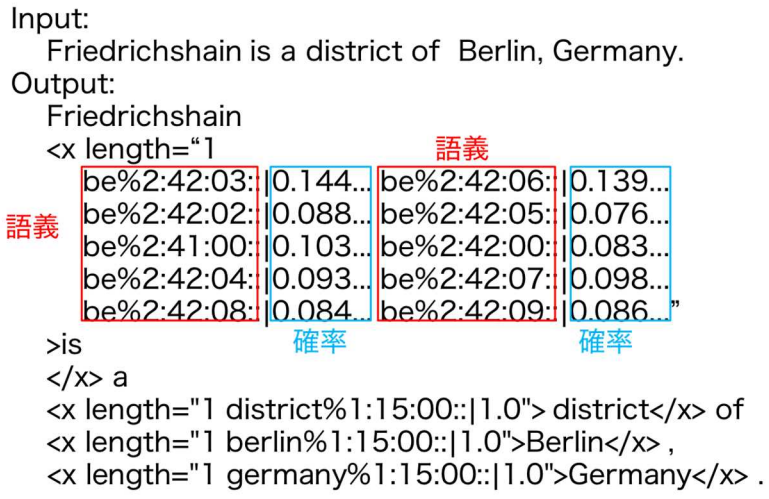

図 2 語義の曖昧性解消

文書に対しては高精度に語義を付与可能な技術であるものの，自然言語に基づかず高々数個の単語から構成されるクエリに対しては，適切に語義を付与することができないと想定

される．従って本研究では，クエリの description を用いて語義の曖昧性解消を行った．クエリの description とは，そのクエリがどのような意図で発行されたのかが自然言語にて記述されており，重要な手がかりとなるため，本課題ではこれを用いた (DescWSD)．ただし，一般的な Web 検索ではこのような description の存在を仮定することはできないため，文法情報を用いず単語間の共起関係のみから語義の曖昧性解消を行う手法 (GrammarlessWSD) の提案も行った．評価実験では，表 1 の通り，DescWSD の方がより高精度となったため，今後 GrammarlessWSD の改良を行う予定である．

表 1 Description を用いた曖昧性解消

	nDCG@20	MAP
DescWSD	.341	.397
GrammarlessWSD	.289	.363

上記によって文書とクエリそれぞれに対して語義が付与されれば，両者の比較を行う．具体的には，同一の語義がそれぞれに高い確率で付与されているほど局所的な意味の類似度が高くなると考えられる．その際，IMS によってもっとも高い確率を付与された語義のみを用いる手法 (BestSense) とすべての語義を用いる AllSense を提案し，実験から BestSense がより高精度であることが示された．既存の特徴量のみを用いた手法 (Common) と比較して統計的有意に検索精度の工場が確認されたことより，語義の曖昧性解消によって適切に局所的な意味を取り込んだモデルの構築に成功し，研究テーマ (1) は達成された．

表 2 実験結果

	nDCG@20	MAP
[Baseline]		
BM25 [18]	.267	.282
Common	.334	.386
[WS score]		
BestSense	.341*	.397**
AllSense	.333	.397**
[DR score]		
DiffDR	.338	.396**
DistDR	.335	.386**
CosDR	.341	.397**
BestSense+CosDR	.349**	.390*

研究テーマ (2)

クエリに対して付与された検索タスクを特徴量として付与したモデルを構築した場合に，現時点では精度向上に結びつけることができておらず，より詳細な分析を行うため，検索タスクごとにどのような文書，さらにはどのような特徴を持つ箇所が適合情報となるのかを分析した．その結果，検

索タスクごとに粒度や構造の観点において適合情報の性質が大きく異なるという結果が得られた．これらの結果を踏まえて，文書検索よりもより直接的にユーザの求める情報を提示する要約型情報検索においては検索タスクを用いることでより有益な検索結果を提示することができるということが示唆された．従って，研究テーマ (2) は現時点では達成できていないものの，今後達成に向けての手がかりを得た．

研究テーマ (3)
本課題の設定では，クエリは訓練クエリとテストクエリに分割されており，つまり，クエリからは適合文書の特徴を類推することができない．これはつまり，既存研究では文書の分散表現を特徴量として追加することで適合文書の特定に利用しているが，本課題ではクエリの持つ分散表現 (query2vec) の特徴と文書の持つ分散表現の特徴量 (doc2vec) の類似度から推定する必要があるということである．二つのベクトルが同一の情報を持つ場合にはベクトルの差は零ベクトルとなることが想定されるため，DiffDR では非類似度として doc2vec と query2vec の差を特徴量として追加した．また，差ではなく距離として非類似度を計測した手法 DostDR や，両ベクトルのコサイン類似度を用いる CosDR の提案も行った．これらの結果，表 2 の通り，CosDR が最も高精度を示した．また，BestSense と CosDR は独立した特徴量であるため両者をかけ合わせた手法である BestSense+CosDR はすべての手法の精度を上回り，統計的にも有意に検索精度の向上が確認された．このことから，研究テーマ (3) は達成できた．

3. 今後の展開

本課題では文書単位の情報検索に着目して取り組んできた．今後の展開としては，文書中からユーザが求める情報そのものを抽出して提示する要約型情報検索への適応を目指す．本研究において完全に解決し切れなかった課題である検索タスクを考慮したランキング学習に関して，要約型情報検索においては検索タスクごとに適合情報の性質が大きく異なるという知見が得られているため，次世代の検索システムのパラダイムとして見込まれる要約型情報検索においては検索タスクを考慮したランキング学習はより実用的な技術となることが期待される．

また，本研究では，ユーザの状況を考慮していなかったため，ユーザの気分や緊急度に基づいて検索結果の構築方針を動的に調整する機構の実現を目指す．

4. 自己評価

本研究で掲げた三つの課題のうち，課題 (1) と課題 (3) は解決でき，課題 (2) においては解決に向けての取り掛かりを発見できたことから，概ね目的を達成できたと考える．本課題はすべて研究代表者が主体的に取り組み，また，適宜専門家への助言を求めて取り組んだため，研究実施体制は適正であった．また，研究費はルールに則り執行した．本研究の成果は条件的には極めて高性能な結果を示したため，ランキング学習における一つの方向性を示せたと考える．今後取り組む予定の要約型情報検索への適応も実現すれば，情報検索のパラダイムを変更できるポテンシャルを持つと考える．なお，本研究では，これまで十分に検討されてこなかった，自然文法に基づかないクエリに対しても適切な自然言語処理 (語義の曖昧性解消) を行うことを目指したという点において極めて挑戦的かつ独創的であった．

5. 主な研究成果リスト

(1) 論文 (原著論文) 発表

なし

(2) 特許出願

研究期間累積件数:0 件

(3) その他の成果 (主要な学会発表、受賞、著作物、プレスリリース等)

1. Atsushi Keyaki, Kazunari Sugiyama, Min-Yen Kan, Jun Miyazaki: “Preliminary Experiments on Semantic Tagged Information Retrieval”, Young Researcher Association for NLP Studies, 2017. (査読なし国内会議)

2. Atsushi Keyaki and Jun Miyazaki: “Analysis of Relevant Text Fragments on Different Search Task Types”, in Proceedings of the 14th Asia Information Retrieval Societies Conference (AIRS), 2018. (査読付き国際会議)

3. 欅惇志, 宮崎純: “要約型情報検索における適合情報のクエリのタスクごとの複雑性調査”, 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM 2019), 2019. (査読なし国内会議)