サーベイ論文作成支援のための引用論文推薦

飯沼　俊平†　　難波　英嗣†　　竹澤　寿幸†

† 広島市立大学大学院 情報科学研究科　〒 731–3194 広島市安佐南区大塚東三丁目 4 番 1 号
E-mail: †{iinuma,nanba,takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし　学術情報量が爆発的に増加している今日，研究者が関連論文全てに目を通し，利用することが困難になっている．このような状況から，特定の研究分野に関連したサーベイ論文や専門書籍の必要性が高まっている．先行研究では，論文間の引用関係に着目し，引用論文データベースからサーベイ論文を自動的に検出する手法が提案されている．しかし，研究者が知りたい分野のサーベイ論文を見つけても，その論文が何年も前に執筆されたものであった場合，最新の研究動向を把握することができない．本研究では，既存サーベイ論文をもとに，そこに加えるべき新しい論文の推薦を試みる．

キーワード　情報推薦，学術論文，サーベイ論文作成支援

1. はじめに

研究者数の増加，学問分野の専門分化と共に学術情報量が爆発的に増加している今日，研究者が入手できる論文の量も増える一方で，人間の処理能力の限界から，入手した論文全てに目を通し利用することが益々困難になっている．このような状況にあって，特定の研究分野に関連したサーベイ論文や専門書籍の必要性は高まる一方である．例えば，医学分野ではコクラン共同計画のもと，サーベイ論文の作成と更新が定期的に行われている．しかし，それ以外の研究領域では同様の仕組みが整備されていないため，研究者が知りたい分野のサーベイ論文を見つけても，その論文が何年も前に執筆されていて，最新の研究動向を含んでいない，あるいは，該当分野のサーベイ論文自体が存在しないことがある．

このような状況を改善するため，これまでに，研究者が行うサーベイ論文の執筆を支援するシステムを開発する研究や [6], [7]，サーベイ論文そのものを複数の論文から自動生成する研究 [5] が行われてきている．Nanba らは，論文間の引用関係に着目し，引用論文データベースからサーベイ論文を自動的に検出する手法を提案している [7]．この手法により，もし研究者が必要とする分野のサーベイ論文が検出されれば，その分野の効率的なサーベイが可能となるが，上述のとおり，検出されたサーベイが最新の研究動向を含んでいるとは限らない．Mohammad らは，論文間の引用関係などを利用して，サーベイ論文の自動作成を試みている．この手法は，要約対象となる複数のテキストから重要文を抽出し，抽出された文間の結束性を考慮して並べることにより要約作成を行う，一般的な複数テキスト要約の手法に基づくものである．しかしながら，現状では数文程度の非常に短い分野の概要を生成する程度にとどまっている．

我々は，Nanba らの研究を発展させ，検出されたサーベイ論文に新しい研究を追加することにより，最新の研究動向を含んだサーベイ論文の自動作成を目指している．Mohammad らの研究に代表される従来の複数テキスト要約技術では，要約対象となる複数のテキストから重要と思われる個所を抽出し，それらを並べて出力するという手法が一般的であった．しかし，この方法では，サーベイ論文のような非常に長い文書を自動生成するのは困難である．その理由は，現在の自然言語処理技術は，文間の結束性などの局所的な情報を考慮してボトムアップに文脈を生成することは可能でも，大域的な観点から非常に長い文書の文脈をするまでには至っていないからである．これに対し，我々が目指している手法では，大域的な観点についてはサーベイ論文の著者の視点を利用することで文書全体の構造を維持しつつ，局所的な文脈解析および文脈生成技術を用いて既存のサーベイ論文に最新研究の情報を追加することで，従来の複数テキスト要約手法よりも非常に長い文書の生成が可能となると考えている．

本研究では，上述したサーベイ論文の自動作成に必要な要素技術である，既存サーベイ論文に追加すべき新しい論文の検索を試みる．すなわち，既存サーベイ論文や専門図書をもとに，そこに追加すべき新しい論文を検索する．サーベイ論文の自動作成には至らずとも，新しい関連論文を示すことで，さまざまな研究分野のサーベイの効率化につながると考えられる．

2. 関連研究

サーベイ論文は，複数論文の要約と捉えることができ，実際に，一般的な複数テキスト要約手法を用いてサーベイ論文を自動作成する試みがなされている [5]．また，特定の論文に対し関連研究の要約を自動生成するなど [2]，学術論文に特化した複数テキスト要約手法が提案されている．Hoang らは，トピック木を要約を構成する際の補助的なデータとして用い，複数論文から関連研究の要約を自動作成している．Jaidka らは，学術論文中の関連研究の要約が持つ，先行研究との比較や研究の位置づけを示すなどの修辞的な役割に着目し，複数論文要約のための枠組みを考案している [3]．要約を作成する際，どのような観点から情報をまとめるか，ということが非常に重要であり，我々は既存サーベイ論文が持つ構造を，要約を構成する際の手がかりとして用いることができると考えている．

サーベイ論文自動作成には，自動要約技術に加え，要約対象とする論文検索技術が必要である．研究者に対し，過去に執筆した論文やそれらの引用関係をもとに学術論文を推薦する研究が行われている．Sugiyama らは，協調フィルタリングを学術論文の引用ネットワークに適用して潜在的引用論文（potential citation paper; 明示的な引用はなされていないが，関係性が高い論文）を検出し，引用ネットワークに加えて利用することで，高い推薦精度が得られることを示している [9]．He らは，ノンパラメトリックな確率モデルにもとづく引用文献推薦システムを CiteSeer(注1) にて構築しており [1]，論文のタイトル，概要，引用が必要なコンテキストを入力とし，その論文が全体として引用すべき論文と，入力されたコンテキストで引用すべき論文を推薦する．コンテキストを入力として論文を検索するという意味では本研究と類似しているが，本研究では，入力では直接言及されていない新しい論文の検索を目的としている．

3. 追加すべき論文の検索

問 題 設 定

特定の研究分野に関するサーベイ論文または専門書籍の本文，および，そこで言及されている論文集合 (書誌情報のみ) を入力として，そこに追加すべき新しい論文集合を出力する．引用論文データベースを検索対象とし，それらから抽出された引用文が利用可能であると仮定する．

3.1 追加すべき論文の候補

既存サーベイ論文に追加すべき論文の候補を取得する手法について，引用関係，入力サーベイや引用論文の書誌情報との類似度に基づく手法を説明する．なお，得られた論文集合は，適合文書の網羅率および集合のサイズを基に評価を行う．

論文間の引用ネットワークは，著者が，無数に存在する関連論文の中から選択した論文から成り，特に関連性の高い論文が取得出来ることが期待できる．ここで，入力サーベイ論文を s，論文集合 P のいずれかを引用している論文集合を $\mathrm{cite}(P)$，論文集合 P のいずれかが引用している被引用論文集合を $\mathrm{ref}(P)$ と表記する．

- $\mathrm{cite}(\mathrm{ref}(\{s\}))$: 入力サーベイ論文が引用している論文集合 $\mathrm{ref}(\{s\})$ の内，いずれかを引用している論文集合．
- $\mathrm{cite}(\mathrm{cite}(\mathrm{ref}(\{s\})))$: $\mathrm{cite}(\mathrm{ref}(\{s\}))$ のいずれかを引用している論文集合．
- $\mathrm{ref}(\mathrm{cite}(\mathrm{ref}(\{s\})))$: $\mathrm{cite}(\mathrm{ref}(\{s\}))$ のいずれかが引用している論文集合．すなわち，入力サーベイ論文が引用している論文 $\mathrm{ref}(\{s\})$ のいずれかと共引用の関係にある論文．
- $\mathrm{cite}(\mathrm{ref}(\mathrm{cite}(\mathrm{ref}(\{s\}))))$: $\mathrm{ref}(\mathrm{cite}(\mathrm{ref}(\{s\})))$ のいずれかを引用している論文集合．

ある論文に関連性の高い論文が多数存在していても，実際に引用される論文数はページ数の制限などにより，非常に限られたものになる．入力された本文や，そこで引用されている論文の書誌情報との類似性を基に候補を取得する．なお，テキスト間の類似性尺度には，Okapi BM25 [8] を用いる．

- $\mathrm{text\text{-}sim}(\mathrm{text}(s), N)$: 入力された本文 $\mathrm{text}(s)$ と，全文との類似度が高い論文を上位 $|\mathrm{ref}(\{s\})| \times N$ 件取得する．ここでは，$\mathrm{text}(s)$ に高頻度で出現する名詞句，上位 50 件を用いる．
- $\mathrm{meta\text{-}sim}(\mathrm{ref}(\{s\}), N)$: 入力が引用している論文と，著者名，タイトル，出典 (会議名，ジャーナル名) が類似する論文を上位 $|\mathrm{ref}(\{s\})| \times N$ 件取得する．ここでは，$\mathrm{ref}(\{s\})$ に含まれる論文の著者名，タイトル，出典に出現する単語を頻度で重み付けして，フィールドごとに類似度を算出し，その和を用いる．
- $\mathrm{citance\text{-}sim}(\mathrm{citances}(s), N)$: 入力に含まれる引用文 $\mathrm{citances}(s)$ と類似する引用文で言及されている論文を，$p \in \mathrm{ref}(\{s\})$ ごとに上位 N 件ずつ取得する．

3.2 ランキング

本節では，3.1 節で説明した手法により得られた論文をランキングし，追加すべき論文を決定する手法について述べる．3.1 節で説明したように，ある論文を新しいサーベイ論文に含めるか否かは，引用関係や内容の類似度，メタ情報 (e.g. 著者名，会議名) など，様々な観点から評価することができる．本研究で用いる指標を次に示す．

- text-sim: 入力サーベイ論文の本文と，対象論文の本文 (全文) との類似度．候補の取得に用いる text-sim() と同様．
- citance-sim: 入力サーベイ論文に含まれる引用文と，引用文データベースに存在する対象論文に関する引用文との類似度の最大値．
- cite-count: 対象論文の被引用数．
- cocite-count: 対象論文が，入力サーベイ論文で言及されている論文と共引用された回数．被引用数から分かる重要度だけでなく，入力サーベイとの関連度を表すことが期待できる．
- meta-sim (author-sim，title-sim，venue-sim): 候補の取得に用いる meta-sim() では著者名，タイトル，出典の類似度の和を用いていたが，ここでは，それぞれを一つの指標として扱う．著者や会議，ジャーナルのオーソリティを表すことが期待できる．

以上で説明した指標で得られたスコアの重み付き線形和を対象論文のスコアとし，ランキングを行う．本研究では，Ranking SVM を用いて指標の重みを学習する．

4. 実　　験

4.1 実 験 方 法

改訂版が出版されている専門書籍を用いて，旧版をもとに，新版で追加された論文を検索することで手法の有効性を検証する．例えば，特定分野に関する専門書籍の第 1 版と，そこで引用されている文献集合をシステムへの入力とし，新しい論文を検索する．第 2 版で追加された論文を適合文書とし，検索結果を評価する．ただし，専門書すべてを入力とすると，内容が広範囲になってしまうため，特定の分野に関する一章を入力とし，対応する新版の章で追加された論文を適合文書とする．

4.2 データセット

表 1 に評価データとして用いる専門書籍を示す．書籍は OCR によりテキスト化し，人手でテキストと引用文献リストが章単

(注1)：http://citeseerx.ist.psu.edu/

位で利用可能な xml 形式に整形した．

表 1　評価に用いる専門書籍（括弧内の数値は出版年を表す．）

書籍タイトル	旧版	新版
“Information Retrieval”(注2)	1st ed. (1998)	2nd ed. (2004)
“Modern Information Retrieval”(注3)	1st ed. (1999)	2nd ed. (2011)
“Speech and Language Processing”(注4)	1st ed. (2000)	2nd ed. (2009)
“Modern Operating Systems”(注5)	2nd ed. (2001)	3rd ed. (2007)

4.1 節で説明したように，章単位での実験を行う．例えば，“Modern Information Retrieval” 第 1 版の 3 章（Retrieval Evaluation）および，そこで参照している文献リストをシステムへの入力とし，第 2 版の 4 章（Retrieval Evaluation）で新たに追加された論文を検索する．このように，タイトルが同一，または類似する旧版と新版の章のペアを 36 組作成し，評価データとした．なお，新版で追加された文献には Web ページや技術仕様書なども含まれているが，本研究では学術論文のみを検索対象とする．適合文書数はトピック平均 36.7 件である．検索対象としては，CiteSeer の全文データ 2,133,683 件とそこから抽出した引用文 49,440,114 件を補助的なデータとして用いる．適合文書のうち CiteSeer に含まれていない論文は，人手で Google Scholar で検索し，データベースに追加した．なお，引用文は ParsCit(注6) によって抽出された引用箇所のうち，文中に “Nanba 2010” や “[2]” のように，引用が明示されている文のみを用いる．

4.3 比較手法

提案手法の有効性を検証するために，次に示す手法でランキングを行い，結果を比較する．

- all-feature (提案手法): 3.2 節で説明した指標すべてを用いてスコアを算出する．ここでは，各スコアを $[0, 1]$ に正規化し，重みは設定せずに足し合わせる．
- RankSVM (提案手法): 指標の重みを Ranking SVM [4] を用いて学習し，スコア算出の際に用いる．実験の際は，表 1 のデータを用いて 4 分割交叉検定を行う．訓練データに関しては，取得した候補のうち正解以外はすべて負例として Pairwise 学習を行う．
- PageRank: 得られた論文の候補に対して PageRank を適用し，スコアを算出する．
- HITS: PageRank と同様，リンク構造からスコアを算出する手法である．ランキングにはオーソリティスコアを用いる．

この他，指標を単体で用いた場合とも比較を行う．評価尺度には，再現率および MRR (Mean reciprocal rank) を用いて評価する．

(注2)：by David A. Grossman and Ophir Frieder. Springer.
(注3)：by Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Addison Wesley.
(注4)：by Daniel Jurafsky and James H. Martin. Prentice Hall
(注5)：by Andrew S. Tanenbaum. Prentice Hall
(注6)：http://aye.comp.nus.edu.sg/parsCit/

5. 結　　果

5.1 追加すべき論文の候補

表 2 に，取得した論文集合に関して，適合文書の網羅率およびサイズを示す．引用関係で論文を取得した場合，単体では ref(cite(ref($\{s\}$))) (入力サーベイ論文が引用している論文のいずれかと共引用の関係にある論文) が，比較的小さいサイズで高い網羅率を得ている．cite(ref(cite(ref($\{s\}$)))) (入力サーベイ論文が引用している論文と共引用関係にある論文のいずれかを引用している論文集合) の網羅率が 0.618 と最も高いが，適合文書数が 36.7 件であることを考慮すると，候補のサイズが 76,303 件と非常に大きい．

本文や書誌情報の類似を用いても 6 割程度収集できるが，引用関係を複数回辿った場合と同様，サイズが大きくなってしまう (text-sim(text(s), 300)，meta-sim(ref($\{s\}$), 300))．引用文の類似性を利用した場合も同様で (citance-sim)，取得件数を増加させても，網羅率に大きな改善は見られない．サイズが 10,000 件前後に収まるように，和集合をとって網羅率を調査した結果 ref(cite(ref($\{s\}$))) $\cup$ meta-sim(ref($\{s\}$), 100) (共引用の関係にある論文と書誌情報が似ている論文) が最も高い網羅率を得た．よって，以降のランキングの実験はこれらの論文集合を対象に行なった．

表 2　候補セットの比較（サイズはトピック平均）

候補セット	網羅率	サイズ
cite(ref($\{s\}$))	0.329	3,567
ref(cite(ref($\{s\}$)))	0.569	8,248
cite(cite(ref($\{s\}$)))	0.462	12,736
cite(ref(cite(ref($\{s\}$))))	0.618	76,303
text-sim(text(s), 100)	0.477	6,142
text-sim(text(s), 200)	0.578	12,272
text-sim(text(s), 300)	0.627	18,389
meta-sim(ref($\{s\}$), 100)	0.468	5,265
meta-sim(ref($\{s\}$), 200)	0.557	10,575
meta-sim(ref($\{s\}$), 300)	0.598	15,895
citance-sim(citances(s), 100)	0.262	1,938
citance-sim(citances(s), 200)	0.339	3,626
citance-sim(citances(s), 300)	0.387	5,210
ref(cite(ref($\{s\}$))) $\cup$ cite(ref($\{s\}$))	0.663	10,979
ref(cite(ref($\{s\}$))) $\cup$ text-sim(text(s), 100)	0.726	13,638
ref(cite(ref($\{s\}$))) $\cup$ meta-sim(ref($\{s\}$), 100)	0.746	12,795

5.2 ランキング

5.1 節で比較的高い網羅率を得た論文集合 ref(cite(ref($\{s\}$))) $\cup$ meta-sim(ref($\{s\}$), 100) を対象にランキングを行なった．上位 k 件の再現率を図 1，MRR による評価結果を表 3 に示す．なお，PageRank，HITS，引用数により算出したスコアをもとにランキングした結果，上位 300 件での再現率がそれぞれ，0.00491，0.0279，0.0694 と非常に低い値であったため，図 1 からは省略している．図 1 より，すべての指標を用いた RankSVM が最も高い再現率を得ることができた．一方 MRR による評価では，

指標の重みを考慮しない all-feature により最も高い値を得た．指標単体で見ると，共引用数 (cocite-count) が最も有効な指標であることがわかる．

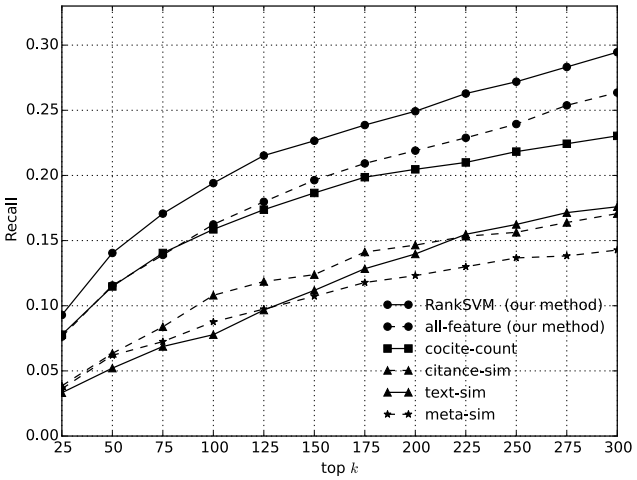

図 1　上位 k 件の再現率

表 3　MRR による評価

ランキング手法	MRR
all-feature (our method)	0.447
RankSVM (our method)	0.385
cocite-count	0.316
meta-sim	0.236
citance-sim	0.193
text-sim	0.106
cite-count	0.047
PageRank	0.022
HITS (authority)	0.018

6. 考　　察

候補を取得する実験では，共引用関係，論文の出典情報が比較的有用であることがわかった．また，ランキングの実験では，特に共引用数が有効であることがわかった．入力サーベイ論文が引用している論文と共引用された回数が多いということは，もともと引用されている論文との関連性が高く，さらに，引用数が多ければ重要であると考えることができる．しかし，すべての論文からの引用数や PageRank，HITS をランキングに利用した場合，再現率，MRR ともに非常に低い値となった．実験に用いた書籍は，旧版が出版されてから新版が出版されるまで，平均約 8 年経過している．そのため，追加すべき論文の候補の数も，適合文書数と比較すると非常に大きくなる．実験では，適合文書の 7 割を網羅するためには 1 万件以上の論文を収集する必要があった．候補が増加と共に，引用数が多いが入力との関連性が低い論文も増加したため，単純な引用関係のみで，入力サーベイ論文との関連性を十分に捉えることが困難になったと考えられる．

最も高い再現率を得た RankSVM でも，上位 100 件で 0.2 程度であり，実際に取得できた適合文書は非常に少ない．しかし，取得した論文を確認したところ，非常に関連性の高い論文を取得できていることがわかった．表 4，5 に取得した論文の例を示す．いずれも，新版には追加されなかったが，入力と関連性が高く被引用数も多い論文を含んでおり，提案手法が有効であることがわかる．ランキングを困難にしている原因は，関連論文の多さにあると考えている．特に，複数の分野に関わりのあるテーマの場合は難しくなる．たとえば，表 5 の “Deadlock” は，複数のプロセスが動作するシステムで発生しうる問題であり，オペレーティングシステムだけでなく，データベースやプログラミング言語処理系など，関係のある分野が多数存在する．

表 4　“Distributed Information Retrieval” (In Information Retrieval, 1st ed.) を入力として得られた論文

順位	書誌情報	適合性
1	Kleinberg. 1999. “Authoritative sources in a hyperlinked environment.” J. ACM	**true**
2	Xu. 1999. “Cluster-based language models for distributed retrieval.” SIGIR ’99.	**false**
3	Callan. 2001. “Query-based sampling of text databases.” ACM Trans. Inf. Syst.	**false**

表 5　“Deadlocks” (In Modern Operating Systems, 2nd ed.) を入力として得られた論文

順位	書誌情報	適合性
1	Subramonian. 2004. “Middleware specialization for memory-constrained networked embedded systems.” RTAS ’04.	**false**
2	Azougagh. 2005. “Resource co-allocation: a complementary technique that enhances performance in Grid computing environment.” ICPADS ’05.	**false**
3	Damron. 2006. “Hybrid transactional memory.” ASPLOS XII.	**false**

7. おわりに

本研究では，既存サーベイをもとに，そこに追加すべき新しい論文の検索を試みた．実験では，特に共引用関係が重要な指標であることがわかり，リンク構造を利用した PageRank や HITS などと比べ，高い再現率を得ることが出来た．しかし，共引用数をもとにした場合，入力で言及されている論文が古くなるにつれて単位時間あたりの被引用数は減少するため，最新の論文を取得することは難しくなる．今回，ランキングの対象とした論文集合にも，一度も共引用されていない論文が含まれていた．一度にすべての論文を取得しようとするのではなく，検索対象を年代で区切り，見つかった論文を入力に追加・再検索を繰り返すことで，改善されるのではないかと考えている．

参 考 文 献

[1] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware Citation Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 421–430, 2010.

[2] Cong Duy Vu Hoang and Min-Yen Kan. Towards Automated Related Work Summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pp. 427–435, 2010.

[3] Kokil Jaidka, Christopher Khoo, and Jin-Cheon Na. Deconstructing Human Literature Reviews - A Framework for Multi-Document Summarization. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pp. 125–135, August 2013.

[4] Thorsten Joachims. Training Linear SVMs in Linear Time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, p. 217, New York, New York, USA, 2006. ACM Press.

[5] Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. Using Citations to Generate Surveys of Scientific Paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pp. 584–592, 2009.

[6] Hidetsugu Nanba, Atsushi Fujii, Mokoto Iwayama, and Taiichi Hashimoto. Overview of the Patent Mining Task at the NTCIR-8 Workshop. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp. 293–302, 2010.

[7] Hidetsugu Nanba and Manabu Okumura. Automatic Detection of Survey Articles. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'05, pp. 391–401, 2005.

[8] Stephen Robertson and Stephen Walker. Some Simple Effective Approximations to the 2-poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pp. 232–241, 1994.

[9] Kazunari Sugiyama and Min-Yen Kan. Exploiting Potential Citation Papers in Scholarly Paper Recommendation. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pp. 153–162, 2013.