新情報の追加によるサーベイ論文の作成支援

飯沼 俊平 † 難波 英嗣 ‡ 竹澤 寿幸 ‡

† 広島市立大学 情報科学部
‡ 広島市立大学大学院 情報科学研究科

{iinuma,nanba,takezawa}@ls.info.hiroshima-cu.ac.jp

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．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 [Nanba 05]．しかし，この手法により検出されたサーベイ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．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．そのための第一歩として，本研究では既存サーベイ論文をもとに，そこでは言及されていない新しい論文の検索を試みる．

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 [Mohammad 09]．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 [Hoang 10]，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 [Jaidka 13]．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．TAC[*1]では，更新情報の要約を自動作成するタスクが提起されている [Dang 08]．ユーザが古い文書集合 B を読んでいると仮定し，新しい文書集合 U から更新情報の要約を作成することがタスクの目的である．ニュース記事に対してさまざまな実験が行われているが，B を既存サーベイ論文，U をそこに追加すべき新しい論文集合とすれば，このタスクでの研究成果をサーベイ論文自動作成に応用可能であると考えられる．

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 [Sugiyama 13]．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[*2]にて構築しており [He 10]，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，我々が目指しているのはサーベイ論文自動作成であり，入力では言及されていない新しい関連論文の検索を目的としている．

[*1] http://www.nist.gov/tac/
[*2] http://citeseerx.ist.psu.edu/

3 追加すべき論文の検索

本節では，既存サーベイ論文に追加すべき論文の検索手法について説明する．引用論文データーベース P を検索対象とし，それらから抽出された引用箇所（citation context; 引用文献に関して言及している箇所）C が利用可能であると仮定する．

入力として，サーベイ論文または専門書籍の特定のトピックに関する一部（節，章）s と，そこで言及されている論文集合 D が与えられ，これらをもとに論文 p $(\in P)$ のスコアを算出する．被引用数が多い論文ほど重要度が高いと考えられるが，特に，d $(\in D)$ との共引用関係を利用することで，特定分野内での重要度を測ることができると考えらる．また，ある論文に関する引用箇所は，他の研究者がその論文に見出した関連性や新規性などの注目すべき点を示しており，引用箇所間の類似性が高ければ，対応する論文対の類似性，関連性が高いと考えられる．上記の考えに基づき，次の 2 つの評価尺度を提案する．なお，テキスト間の類似性尺度として BM25 を用いる．

co-count (co-citation count): d $(\in D)$ との共引用数を被引用半減期[*3]で割った値を co-count1，共引用関係にある D 内の文書数を co-count2．2 つを統合したものを co-count とする．[*4]

l-sim (local similarity): s 中の D に関する引用箇所 C_s と，論文 p に関するすべての引用箇所 c_p $(\in C)$ との類似度を算出し，その最大値を論文 p の l-sim1 とする．引用箇所が他の研究者の観点からの要約であるのに対し，概要は著者自身の観点からの要約である．引用箇所を概要で代用し，C_s と p の概要との類似度を l-sim2 とする．2 つを統合したものを l-sim とする．なお，大域的な類似性も必要であるため，l-sim は，入力 s と p の全文との類似度 g-sim と統合して使用する．

4 実験

3 節で提案した共引用関係に基づく手法 co-count，引用箇所間の類似性に基づく手法 l-sim の有効性を検証するために実験を行った．

[*3] 論文の被引用数の各年毎の累計が，引用された総数の 50% に至るまでの現在から過去への年数

[*4] スコアを統合する際は，1 位のスコアが 1 になるように正規化したうえで足し合わせる．

実験方法

情報分野の専門書籍 4 冊の旧版の一章と，そこで言及されている文献を入力，対応する新版の章で追加された論文を適合文書とみなし，正解データを 43 トピック作成した．CiteSeer の全文テキストを含む書誌情報データ 2,000,380 件を検索対象 P とし，それらから抽出された引用箇所 18,028,360 件を C として用いる．なお，c $(\in C)$ は参照部分から前後 200 文字ずつ抽出されたテキストである．正解データとして用いる専門書籍を表 1 に示す．たとえば，“Modern Information Retrieval” 第 1 版の 3 章（Retrieval Evaluation）および，そこで言及されている文献集合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．適合文書数はトピック平均 17.5 件である．なお，新版の内容を最新の動向情報と仮定するため，検索を行う際，新版の出版年よりも後に発表された論文は検索対象としない．また，引用関係を用いる場合も新版の出版年以前の論文のみを用いる．検索結果は再現率および MRR で評価する．

表 1 正解データとして用いる専門書籍（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”[*5]	1st ed. (1998)	2nd ed. (2004)
“Modern Information Retrieval”[*6]	1st ed. (1999)	2nd ed. (2011)
“Speech and Language Processing”[*7]	1st ed. (2000)	2nd ed. (2009)
“Modern Operating Systems”[*8]	2nd ed. (2001)	3rd ed. (2007)

追加すべき論文の候補

予備実験により D と共引用の関係にある論文集合 D_{co} に適合文書の約 7 割が含まれることがわかっており，それらに対して順位付けを行う．なお，D_{co} の論文数は約 1 万件（トピック平均）である．

比較手法

次に挙げる手法を用いた場合と比較することで，提案手法の有効性を検証する．

[*5] by David A. Grossman and Ophir Frieder. Springer.

[*6] by Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Addison Wesley.

[*7] by Daniel Jurafsky and James H. Martin. Prentice Hall

[*8] by Andrew S. Tanenbaum. Prentice Hall

PageRank: $D_{co} \cup D_{cite}$[*9]の引用ネットワークに対して PageRank を適用しスコアを算出する．

PageRank-bi: 引用関係にある論文対に，双方向にリンクを持たせて，PageRank を適用する．

HITS: PageRank と同様に，リンク構造からスコアを算出する手法である．オーソリティスコアを用いる．

g-count (global citation count): 被引用数を被引用半減期で割った値をスコアとする．

l-count (local citation count): $D_{co} \cup D_{cite}$ からの被引用数を被引用半減期で割った値をスコアとする．

g-sim (global similarity): 入力サーベイ s と論文 p の全文との類似度．

5 結果および考察

共引用関係に基づく手法 co-count と，その他の引用関係に基づく手法による結果を図 1，表 2 に示す．D との共引用回数 co-count1，共引用された D 内の文書数 co-count2 が比較的高い再現率を得ており，MRR は co-count2 が最も高く，$k = 25$ の時 co-count2 の再現率が最も高くなっている．g-count は再現率，MRR 共に最も低く，$D_{co} \cup D_{cite}$ からの被引用数である l-count が比較的高い再現率を得た．すなわち，ある論文集合内での被引用数が多ければ，その分野に限定して特に重要な論文であるといえる．また，共引用関係を利用することで，さらに分野を限定した重要度の評価を行うことができたと考えられる．

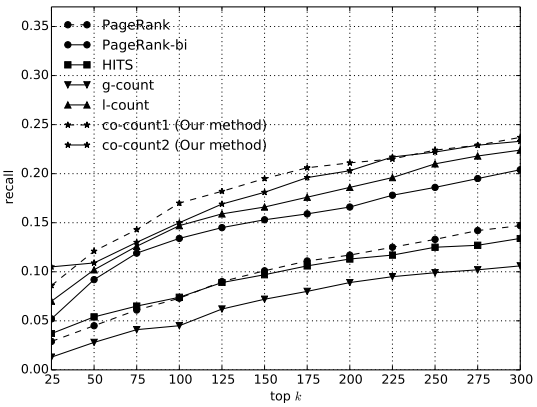

図 1 引用関係に基づく順位付け結果：上位 k 件の再現率

[*9] D_{cite} は D を引用している論文集合

表 2 引用関係に基づく順位付け結果：MRR による評価

順位付け手法	MRR
PageRank	0.083
PageRank-bi	0.142
HITS	0.115
g-count	0.053
l-count	0.152
co-count1 (Our method)	0.196
co-count2 (Our method)	0.239

引用箇所間の類似性に基づく手法 l-sim と，g-sim による結果を図 2，表 3 に示す．入力サーベイと全文との類似度 g-sim に，引用箇所間の類似度 l-sim1，および，引用箇所と概要との類似度 l-sim2 を加えた場合に MRR が最大になっている．引用箇所は，他の研究者が捉えた論文の特徴と考えることができ，これらの類似性を考慮することで，より関連性の高い論文が取得可能であることを示している．

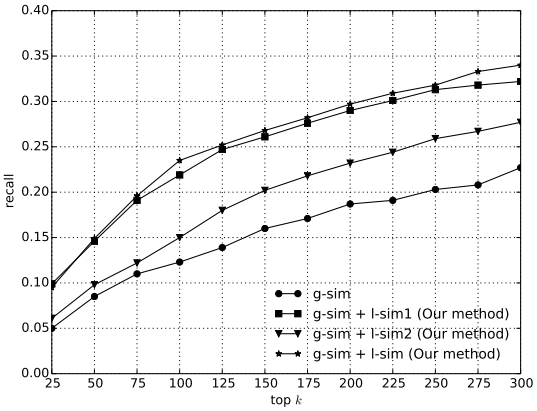

図 2 内容の類似性に基づく順位付け結果：上位 k 件の再現率

表 3 内容の類似性に基づく順位付け結果：MRR による評価

順位付け手法	MRR
g-sim	0.149
g-sim + l-sim1 (Our method)	0.271
g-sim + l-sim2 (Our method)	0.221
g-sim + l-sim (Our method)	0.333

co-count と l-sim を組み合わて実験を行った．なお，l-count + g-sim, PageRank-bi + g-sim は共引用関係および引用箇所を用いない場合の比較手法として挙げた．結果を図 3, 表 4 に示す, l-count + g-sim, PageRank-bi

+ g-sim と比べ，co-count，l-sim を用いた手法での再現率，MRR が高く，提案手法が有効であることがわかった．

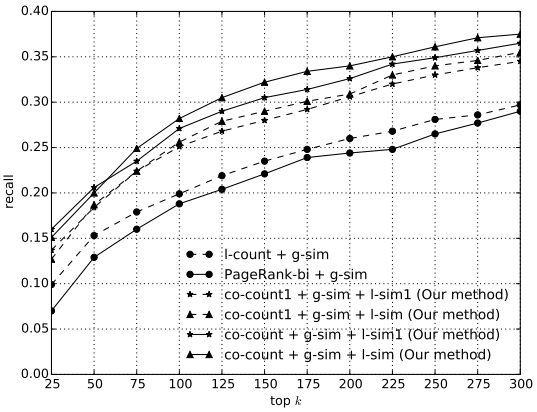

図 3　引用関係および内容の類似性に基づく順位付け結果：上位 k 件の再現率

表 4　引用関係および内容の類似性に基づく順位付け結果：MRR による評価

順位付け手法	MRR
l-count + g-sim	0.216
PageRank-bi + g-sim	0.185
co-count1 + g-sim + l-sim1 (Our method)	0.323
co-count1 + g-sim + l-sim (Our method)	0.356
co-count + g-sim + l-sim1 (Our method)	0.321
co-count + g-sim + l-sim (Our method)	0.373

トピックごとの分析

co-count + g-sim + l-sim を用いた検索結果上位 50 件での再現率をトピックごとに調査した．再現率が高いトピックと低いトピックを比較すると，順位付けに用いた D_{co} の文書数に顕著な差が見られた．統計的構文解析に関するトピックでは，再現率 0.56（正解数：20/36）に対して D_{co} の文書数は 4,372 件であり，インデクシング（情報検索）に関するトピックでは再現率 0.64（正解数：9/14）に対して 7,173 件であった．マルチメディア情報検索に関するトピックでは，再現率 0.07（正解数：1/14）に対し 24,532 件．マルチプロセッサに関するトピックでは，再現率 0.11（正解数：4/35）に対し 20,124 件であり，候補の文書数が多い場合に再現率（上位 50 件の）が低くなる傾向にある．

d_1 ($\in D$) と d_2 ($\in D_{co}$) を共引用している論文 d ($\in D_{cite}$) の中での d_1，d_2 の関係性（引用されている位置など）を考慮し，候補を絞る必要があるといえる．

6　おわりに

既存サーベイ論文をもとに，そこに追加すべき新しい論文の検索を試みた．共引用関係をもとに重要度を評価し，既存サーベイ論文との類似性に加え，引用箇所間の類似性を考慮することで，特に関連性の高い論文を取得可能であることを示した．今後の課題として，追加すべき論文の候補全体にどのような研究課題があり，年度別に見てどのような技術要素が注目を集めているか，など，候補全体に対して詳細な分析を行い，順位付けの手がかりとする必要があると考えている．

参考文献

[Dang 08] Dang, H. T. and Owczarzak, K.: Overview of the TAC 2008 update summarization task, in *Proceedings of Text Analysis Conference*, pp. 1–16 (2008)

[He 10] He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, L.: Context-aware citation recommendation, in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 421–430 (2010)

[Hoang 10] Hoang, C. D. V. and Kan, M.-Y.: Towards automated related work summarization, in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pp. 427–435 (2010)

[Jaidka 13] Jaidka, K., Khoo, C., and Na, J.-C.: Deconstructing Human Literature Reviews – A Framework for Multi-Document Summarization, in *Proceedings of the 14th European Workshop on Natural Language Generation*, pp. 125–135 (2013)

[Mohammad 09] Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., and Zajic, D.: 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)

[Nanba 05] Nanba, H. and Okumura, M.: 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)

[Sugiyama 13] Sugiyama, K. and Kan, M.-Y.: 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)