新情報の追加によるサーベイ論文の作成支援

A Support System for Updating Survey Articles

飯沼俊平 Shumpei Iinuma　難波英嗣 Hidetsugu Nanba　竹澤寿幸 Toshiyuki Takezawa

広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

1　はじめに

学術情報量が爆発的に増加している今日，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．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 [1]．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　追加すべき論文の検索

本研究では，CiteSeer[1] 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）18,028,360 件を補助的なデータとして用いる．CiteSeer の論文集合を P，引用箇所の集合を C とする．

既存サーベイ論文 s と，そこで言及されている文献集合 D を入力として与えられ，これらをもとに論文 p $(\in P)$ のスコアを算出する．特に，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れる．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: d $(\in D)$ との共引用数を被引用半減期で割った値を 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．[2]

l-sim: 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，C_s と p の概要との類似度を l-sim2，2 つを統合したものを l-sim とする．なお，大域的な類似性も重要であるため，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．

[1] http://citeseerx.ist.psu.edu/

[2] 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．

3　実験

情報分野の専門書籍 4 冊の旧版の一章 [3] 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，実験を行った．なお，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 割が含まれていることがわかっており，それらに対して順位付けを行った．適合文書数は 17.5 件であり，D_{co} の論文数は約 1 万件である（トピック平均）．

結果を図 1 に示す．比較手法として，入力 s と全文との類似度 g-sim，$D_{co} \cup D_{cite}$[4] 内での PageRank スコア，$D_{co} \cup D_{cite}$ 内での被引用数 l-count を組み合わせた手法を挙げた．図から分かるように，再現率を改善することができた．

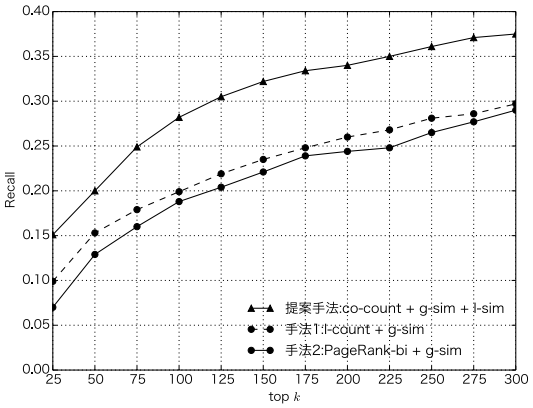

図 1　順位付け結果：上位 k 件の再現率

4　おわりに

既存サーベイ論文中で言及されている論文集合との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て，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できることを示した．

参考文献

[1] Nanba, H. et al. Automatic Detection of Survey Articles. In Proceedings of ECDL 2005, pp.391-401.

発表論文

飯沼 俊平，難波 英嗣，竹澤 寿幸. 新情報の追加によるサーベイ論文の作成支援．言語処理学会 第 20 回年次大会，2014.

[3] “introduction” など，タイトルから分野を特定できないものは対象としない．

[4] D_{cite} は D を引用している論文集合である．