2L-05

リンク解析による情報の断片のランキング

飯塚翔† 湯本高行‡ 新居学‡ 上浦尚武‡

兵庫県立大学工学部† 兵庫県立大学大学院工学研究科‡

1 はじめに

Web 検索エンジンの一般的な利用形態である“10 blue links”では，システムはユーザが入力した検索クエリに対して関連する Web ページへのリンクを複数出力し，ユーザは複数のリンクから Web ページを選択・閲覧する．この利用形態では，ユーザは必要としている情報を満足するまで Web ページの選択・閲覧を繰り返す必要があるため，スマートフォンのように画面が小さく，インターネット接続が低速な環境ではユーザの負担が大きい．このような理由から，複雑な操作なしで素早く情報を得られるシステムの必要性が増している．この課題へのアプローチとして，Web ページ集合に含まれる情報を要約してユーザに提示するという方法が考えられる．この方法は Web ページ集合から情報の断片を抽出するという段階と，情報の断片を整理してユーザに提示するという段階の 2 つに分けられる．情報の断片を整理する段階では，ユーザが読む必要があるテキスト量を減らすという観点から，重要度の高い情報の断片を要約の前方に配置することは有用である．本研究の目的は，情報の断片と Web ページとの関係から情報の断片の重要度を推定することにより，情報の断片のランキングを作成することである．

2 問題設定

本研究では情報の断片の定義として NTCIR12 MobileClick-2 タスク [1] で定義されている iUnit を用いる．iUnit は，情報として有益か，それ以上分けることができないか，という観点で情報の断片の粒度が定義されている．表 1 に iUnit の例を示す．データセットには NTCIR-12 MobileClick-2 タスクの訓練データを使用する．100 種のクエリのそれぞれに対応する iUnit が設定され，それぞれの iUnit にはアノテータが付与した重要度が設定されている．またデータセットとしてクエリを Bing で検索した際の上位 500 件の Web ページも与えられる．本研究では iUnit の重要度を推定することにより重要な iUnit を上位に配置するようなランキングを作成する．

表 1 iUnit の例

クエリ	iUnit
イチロー	本名は鈴木一朗
	所属 元・オリックス・ブルーウェーブ
宇都宮駅 焼き鳥	沖縄料理・焼き鳥琉球酒場 phone 028-678-5628
	沖縄料理・焼き鳥琉球酒場 住所〒321-0953 栃木県宇都宮市東宿郷 2-5-6-2F

Ranking Method for Pieces of Information by Link Analysis
† Sho Iizuka, School of Engineering, University of Hyogo
‡ Takayuki Yumoto, Manabu Nii and Naotake Kamiura, Graduate School of Engineering, University of Hyogo

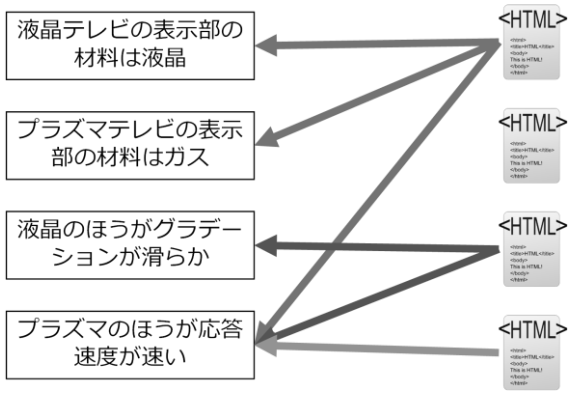

図 1 含意関係を表す二部グラフ

3 提案手法

iUnit は構成する語の数が少ないため，テキストのみを使って重要度を推定することは困難であると考えられる．そこで本研究では iUnit と Web ページの関係を利用する．まず Web ページが iUnit の内容を含意するか推定する．この結果から，Web ページと iUnit を頂点とし，Web ページが iUnit を含意する場合に辺を有する二部グラフを構築する．図 1 に含意関係を表す二部グラフの例を示す．次に，得られた二部グラフでグラフ上の頂点の重要度の推定に有効と考えられる特徴量をリンク解析の手法を用いて算出し，その数値に応じてランキングを作成する．

iUnit と Web ページの含意関係を推定する方法として，iUnit に含まれる語と Web ページに含まれる語の重複度合を用いる方法を利用する．具体的には，iUnit 中の名詞が Web ページに全て含まれるときに含意としてグラフに辺をつくる方法 (ALL)，iUnit 中の名詞が Web ページに 1 つ以上含まれるときに含意としてグラフに辺をつくる方法 (ANY)，iUnit 中の名詞が Web ページに含まれる割合を重みとしてグラフに重み付き辺をつくる方法 (RATE) の 3 種を用いる．ALL と ANY は含意関係を二値で表すのに対し，RATE は含意関係の割合を表すという点が異なる．iUnit と Web ページから名詞のみを取り出すために，日本語タスクでは MeCab [2] と UniDic [3]，英語タスクでは Stanford POS Tagger [4] を用いた．含意関係の推定のために使う品詞の組合せは様々なものを検討したが，最も良い評価値が得られる名詞のみを使う方法を採用した．

次に，得られた二部グラフからグラフ上の頂点の重要度を推定する方法について述べる．iUnit の次数は iUnit を含む Web ページの数であり，多くの Web ページに含まれる iUnit は重要なものであると仮定すれば次数を重要度とする方法 (DEG) が考えられる．また，二部グラフに対して PageRank [5] を

適用し PageRank スコアを重要度とする方法 (PR) が考えられる．PageRank スコアは以下の式により算出される．$adj(v)$は頂点vに隣接する頂点集合，$deg(u)$は頂点uの次数，Nは頂点数である．dはダンピングファクタと呼ばれる係数であり今回は 0.85 に設定する．

$$p(v) \leftarrow \frac{(1-d)}{N} + d \sum_{u \in adj(v)} \frac{1}{deg(u)} p(u) \quad (1)$$

また，二部グラフに対して HITS [6] を適用し，iUnit の権威度スコアを重要度とする方法 (HITS) が考えられる．HITS は以下の式により算出される．$h(v)$はハブスコア，$a(v)$は権威度スコアと呼ばれる．

$$h(v) \leftarrow \sum_{u \in adj(v)} a(u) \quad (2)$$

$$a(v) \leftarrow \sum_{u \in adj(v)} h(u) \quad (3)$$

含意関係を推定する方法に RATE を使用すると重み付きグラフが得られるが，この場合には次数を用いる方法は適用せず，重み付きグラフに対する PageRank, HITS は Mihalcea による手法 [7] を用いた．

4 評価実験

提案手法を用いて作成したランキングを Q-measure [8] によって評価した．Q-measure は重要度の低いアイテムが上位に出現するようなランキングにペナルティを科す評価指標であり，アイテムがアノテータによって付与された重要度によって降順に並ぶ理想的なランキングのとき 1 となる．Q-measure の算出方法を以下に示す．

$$Q = \frac{1}{R} \sum_{r=1}^{M} IsRel(u_r) \frac{\sum_{r'=1}^{r} (\beta \cdot GG(u_{r'}) + IsRel(u_{r'}))}{\sum_{r'=1}^{r} (\beta \cdot GG(u_{r'}^{*}) + 1)} \quad (4)$$

ここでMはランキング長，$GG(u_r)$はシステムの出力したランキングでr位の iUnit の重要度，$GG(u_r^{*})$は理想的なランキングでr位の iUnit の重要度，βは定数であり本研究では 1 としている．$IsRel(u)$は iUnit u がクエリに関係するとき 1，関係しないとき 0 となる関数である．Rはランキング中に含まれている関係する iUnit の数である．

日本語タスクの結果を表 2 に，英語タスクの結果を表 3 に示す．日本語タスクの場合，最も評価値が高い ALL+PR を用いる組合せでは，ベースライン手法であるランダムなランキング (0.773)，ユニグラム言語モデルを使った手法 (0.790) と比較して高い評価値が得られている．含意関係の推定方法を比較すると，最も評価値が高いのは ALL であり，残りは RATE, ANY の順となった．英語タスクの場合，最も評価値が高い ANY+HITS を用いる組合せでは，ベースライン手法であるランダムなランキング (0.803) と比較すると高い評価値が得られているが，ユニグラム言語モデルを使った手法 (0.877) とは同程度の評価値である．含意関係の推定方法を比較すると，最も評価値が高いのは ANY であり，残りは RATE, ALL の順となった．

表 2　日本語タスクにおける Q-measure

	ALL	ANY	RATE
DEG	0.831	0.760	-
PR	0.834	0.762	0.810
HITS	0.823	0.758	0.807

表 3　英語タスクにおける Q-measure

	ALL	ANY	RATE
DEG	0.825	0.879	-
PR	0.823	0.879	0.861
HITS	0.828	0.881	0.857

5 考察

含意関係の推定方法と Q-measure による評価値の関係に着目すると，日本語タスクと英語タスクで評価値の順位が逆転している．この原因として，ALL は iUnit に名詞が多いとき含意と推定されにくく，ANY は iUnit に名詞が多いとき含意と推定されやすいというように，ALL と ANY は iUnit の長さに影響を受けやすい指標であることが考えられる．RATE は分母に iUnit に含まれる名詞数を持つため iUnit の長さの影響が緩和されていると考えられ，ALL と ANY の中間の評価値が得られている．

また，リンク解析の手法と Q-measure による評価値の関係に着目すると，DEG と比較して PR や HITS を用いることによる評価値の大きな改善は見られない．PR や HITS を用いることによって，iUnit が良質な Web ページに含まれているときは重要度を大きく上げるといった効果が得られることを期待しているが，本手法において構築した二部グラフではそのような効果が得られていないと考えられる．

今後の展望として，検索エンジンから得られる Web ページの順位を用いて良質な Web ページであるかどうかの情報を組み込むことなどが考えられる．

謝辞 本研究の一部は，平成 27 年度科研費若手研究 (B)「情報の詳細関係に基づく Web ページの組織化」(課題番号 24700097) によるものである．

参考文献

[1] NTCIR-12 MobileClick-2 http://www.mobileclick.org/
[2] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/
[3] 形態素解析辞書 UniDic, http://pj.ninjal.ac.jp/corpus_center/unidic/
[4] Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml
[5] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, Vol.30, pp.107-117 (1998).
[6] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, pp.604-632 (1999).
[7] Mihalcea, R.: Graph-based ranking algorithms for sentence extraction, applied to text summarization, *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (2004).
[8] Sakai, T.: On the reliability of information retrieval metrics based on graded relevance, *Information Processing & Management*, Vol.43, pp.531-548 (2007).