SNS での注目度と子供に人気のある検索クエリを用いた Web ニュース記事の子供向けランキング手法の調査

Examinations of Web News Ranking for Elementary School Children based on Degree of SNS Users' attention and Popular Queries of Children

田中 翔也[*1]，　安藤 一秋[*2]
Shoya TANAKA[*1]，　Kazuaki Ando[*2]
*1 香川大学大学院工学研究科信頼性情報システム工学専攻
*1 Graduate School of Engineering, Kagawa University
*2 香川大学工学部
*2 Faculty of Engineering, Kagawa University
Email: ando@eng.kagawa-u.ac.jp

あらまし：小学校では，新聞を教材として活用する教育である NIE（Newspaper in Education）が行われている．本稿では，NIE で活用できる記事を小学生に推薦する手法についての現状について述べる．
キーワード：NIE，小学生，新聞，推薦

1. はじめに

近年，小学校では新聞を教材として活用する教育（NIE : Newspaper in Education）が実施されている[1]. NIE では，一般的に紙ベースの新聞を利用するが，Web ニュースが利用される機会も増えてきた．しかし，小学生は語彙力と検索力が乏しいため，膨大なニュース記事の中から目的の記事を探し出すことは難しい[2].

そこで本研究では，SNS での注目度と子供に人気のある語，教科書の重要語･単元情報，教師が指定するキーワードなどを総合的に利用して，小学生が興味を持ちやすい，あるいは興味を持ってもらいたい記事を推薦するシステムの構築を目的とする．本稿では，第一段階として，SNS での注目度と子供に人気のある語のみを用いた小学生のための Web ニュース記事のランキング手法について述べる．

2. SNS での注目度と子供に人気にある語を用いたランキング

2.1 基本方針

Web 新聞社サイトの記事アクセスランキングを小学生の記事推薦に利用することも考えられる．しかし，アクセスランキングは，一般読者が記事を読んだ後，興味を持ったか否かについては反映されていない．そこで本研究では SNS に注目する．SNS で注目される記事は多くのユーザが興味をもった記事であり，興味のない記事はポストされにくい傾向がある．また，SNS での注目度が高いニュースは NIE に適した記事の可能性もある．しかし，SNS での注目度に基づくランキングは，あくまで一般ユーザの視点で構成されるため，小学校での NIE に適したランキングとはいえない．そこで，NIE にとって不要な記事はフィルタリングし，教育的に価値のある記事は，上位に位置づける仕組みが必要となる．

本研究では，SNS での注目度と子供に人気のある語，教科書の重要語･単元情報などを総合的に利用して，小学校での NIE で利用できる記事ランキングの生成を目的とするが，まず本稿では，SNS での注目度と子供語を利用したランキングについて検討する．

2.2 子供に人気のある語（子供語）

子供が興味をもつ語が記事に含まれていると子供が興味を持つと仮定する．本稿では，子供向けポータルサイトで公開されるクエリランキングに注目する．以降，本稿では子供向けポータルサイトにおいて人気のあるクエリを子供語と呼ぶ.

2.3 子供語に基づく記事の重要度

記事の重要度は，記事内の単語重要度に子供語の重みを掛けた値の総和として算出する．単語の重要度は，Okapi BM25[3]と TFIDF で計算し，実験により，いずれかを選択する.

記事iに含まれる単語をt_j，その総数をnとするとき，重要度$S(i)$は，式(1)と式(2)で計算される.

$$S(i) = \sum_{j=1}^{n} C(t_j) \cdot BM25(t_j) \qquad (1)$$

$$S(i) = \sum_{j=1}^{n} C(t_j) \cdot TFIDF(t_j) \qquad (2)$$

ここで，$C(t_j)$は子供語の順位情報に基づく重みである．なお，式(2)の TF 値は，記事中の単語総数で正規化した値を利用する.

子供語の集合をKとするとき，$C(t_j)$は，子供語のランキングを基に，式(3)で計算される.

$$C(t_j) = \begin{cases} 1 & t_j \notin K \\ p \cdot \log_r\left(r + rank_{max} + 1 - rank(t_j)\right) & otherwise \end{cases} \qquad (3)$$

ここで，pとrは調整用パラメタである．$rank_{max}$は，利用する子供向けポータルサイトにおける子供語の最大順位であり，$rank(t_j)$はそのサイトにおけるt_jの順位である．式(3)は，子供語の順位が高い程，重

みが大きくなる．もし，t_jが子供語でない場合，$C(t_j)$の値は 1 とする．

2.4 SNS での注目度と子供語に基づく記事の重要度

記事iに対する SNS での注目度$SNS(i)$は，記事に対する Tweet 数と Facebook でのシェア数を基に，各々を正規化した値の和として算出する．

$$SNS(i) = \frac{Twitter(\mathrm{i})}{max_{tweet}} + \frac{Facebook(i)}{max_{facebook}} \quad (4)$$

$Twitter(i)$は，記事iに対する対象期間での Tweet 数，$Facebook(i)$は，記事iに対する対象期間でのおすすめ数である．また，max_{tweet}と$max_{facebook}$は，それぞれ対象期間の記事において，Tweet 数と Facebook のおすすめ数の最大値である．

最終的な記事iの重要度$Score(i)$は，子供語に基づく記事の重要度$S(i)$と$SNS(i)$を用いて，以下の式で計算される．

$$Score(i) = \alpha \cdot SNS(i) \cdot (1-\alpha) \cdot S(i) \quad (5)$$

ここで，αは重みである．最終的には，Score(i)に基づいて，各記事をランキングする．

3. 実験

3.1 正解データの構築

YOMIURI ONLINE と産経ニュースサイト（2014 年 12 月 1 日から 7 日）から収集した記事群に対し，事件や事故などを除く任意の 54 件を抽出した. そして，54 件の記事に対し，NIE に最も適した記事と NIE に利用できる記事を 23 人の被験者に選択してもらった．その後，NIE に最も適した記事と NIE に利用できる記事に対して，得票数に基づくランキングを生成した．本稿では，このランキングを正解データとして利用する．

3.2 目的と方法

実験の目的は，2 つの実験により，式(1)と式(2)の性能評価と式(3)で利用するパラメタpの適正値を調査し，式(5)で生成されるランキングの妥当性を考察する. まず, スピアマンの順位相関係数を利用して，式(5)に基づくランキングと正解データの順位相関性を調査する．パラメタpは，$2^n (1 \le n \le 6)$の範囲で変化させる. 子供語は，Yahoo!きっずとキッズ goo から正解データと同じ期間に収集したものを利用する．式(3)のパラメタrと式(5)のパラメタαは，それぞれ$r = 2$，$\alpha = 0.5$とする．順位に相関性があるだけでなく，被験者が選択した子供向け記事が上位 10 以内に含まれる割合も調査する．

3.3 結果と考察

3.3.1 順位相関係数による比較結果

NIE に最も適した記事を利用した実験において，BM25 と TFIDF に基づくランキングは，それぞれ正解データと弱い正の相関があることがわかった．全体としては，BM25 に基づくランキングは安定しているが，pが大きくなるにつれ，TFIDF に基づくランキングの方が若干よい結果となった．NIE に利用できる記事を用いた実験においては，NIE に最も適した記事を利用した場合と比べ，BM25 に基づくランキングの方が若干良い結果となった．

2 つの結果とも$p = 2\ or\ 4$以降，順位相関係数の値はほとんど変化しなかった．したがって，これらの結果から，BM25 と TFIDF では大きな違いがないといえる．本実験からランキング間に弱い相関が確認できたが，これらの記事が NIE に適した記事であるとはいえない．今後は，NIE に関する学習知識を活用することでランキングを改善する．

3.3.2 上位 10 位に含まれる子供向け記事の割合

NIE に最も適した記事と利用できる記事が提案手法に基づくランキングの上位 10 位に含まれる割合を調査した結果，BM25 と TFIDF に基づく手法に差はなく，$p = 2$以降は，各手法とも 30%程度の記事が含まれていた．この結果に対しても，NIE に関する学習知識を活用することで，改善できると考える．

3.3.3 パラメタ p の適正値の考察

5.3.1 と 5.3.2 の結果を基に, パラメタpの適正値を考察する．NIE に最も良い記事の結果からは，$p = 2$以降，BM25 と TFIDF の両手法において順位相関係数の値はほとんど変化していなかった．NIE に利用できる記事の結果からは，TFIDF は，$p = 2$以降，BM25 は, $p = 4$以降, 値はほとんど変化しなかった．また, $p = 2$以降, BM25 と TFIDF の両手法において，上位 10 位に含まれる記事の割合の平均値は変化しなかった．以上の結果からパラメタpの適正値について考察する．パラメタpに基づく式(3)は，記事中の単語重要度に対する重みである．pが大きくなると子供語の有無で記事の重要度が決まることになるため, 記事中の単語重要度が無視されることになる．したがって，本稿では，$p = 2$ を適正値と定める．

4. おわりに

本稿では，SNS での注目度と子供語を用いた Web ニュース記事のランキング手法について述べた．また生成したランキング上位 10 位に, NIE に利用できる記事が 30%程度含まれることを確認した. 今後は，学習知識（教科書の単元情報，重要語，教師の指定キーワードなど）を利用して，ランキング精度を向上させる必要がある．また，記事推薦システムを構築し，有用性の確認を行う予定である．

謝辞

本研究の一部は JSPS 科研費 25350335 の助成を受けて実施した．

参考文献

[1] NIE 教育に新聞を，http://nie.jp/

[2] 坪井他：“小学生向け NIE を対象とした Web 新聞記事の推薦.”，情報処理学会研究報告．コンピュータと教育研究会報告, pp.18: 1-5(2013)

[3] S. Robertson et al., “Okapi/Keenbow at TREC-8”, Proc. of the 8th Text Retrieval Conference, 1999.