D-025

更新履歴による注目度を利用した番組検索結果のリランキング

Re-ranking Method for Relevant Program Retrieval using Measure of Attention based on Wikipedia Revision History

後藤淳† 宮崎勝† 田中英輝† 相澤彰子‡
Jun GOTO Masaru MIYAZAKI Hideki TANAKA Akiko AIZAWA

1. はじめに

放送済みの番組を蓄積し、インターネットにより番組を配信するサービスが増加している[1]。我々は、番組概要や字幕のテキストを利用し、番組アーカイブから関連する番組を検索するシステムを開発した[2]。システムでは、番組アーカイブ内の番組概要の n-gram の頻度や、固有表現の情報を利用している。しかし限られたアーカイブ内の情報だけでは、世の中で注目されている事象が何であるのかを考慮することができないため、ユーザが興味をもつ時事性のある関連番組を検索結果の上位に提示できないことがある。そこで、番組アーカイブ外の情報を用いて、注目されている語（注目語）を特定し、その有無により番組検索の結果をリランキングすることを目指す。

注目語を判断する代表的な方法に Web 空間の単語の頻度を利用する方法やサーチエンジンのクエリ語を利用する手法がある[3][4]。これらは、商用検索エンジンが提供する API を利用し、単語のヒット数やクエリ語の頻度を取得している。しかし、番組検索のための単語の注目度に利用するには、頻繁にデータベース中の大量の単語のヒット数を取得する必要があり、各エンジンの回数制限などから、API の利用は実用的には難しい。また多くのショッピングサイトではユーザログを利用した商品の推薦が行われている[5]。このようなユーザログを利用し、高頻度にアクセスされた商品の説明文から注目語を取得することが考えられる。ただし、このような方法が適用可能なのは、一定規模のユーザログが得られるサービスに限られる。

本研究では、誰もが容易に利用できる Wikipedia の記事の変更頻度を注目度として用い、関連番組検索の結果をリランキングする手法を提案する。Wikipedia の変更履歴を利用する関連研究には、編集された複数の記事を比較することで文圧縮のための学習コーパスを自動獲得する研究[6]や、継続的な変更履歴を利用した記事の信頼モデルに関する研究[7]があるが、今回は、変更された内容については考慮せず、ある期間の変更の頻度のみを注目度として利用可能かどうかを検討する。

2. 変更履歴に基づく注目度

2.1 Wikipedia 変更履歴

Wikipedia は世界中のユーザが自由に編集できるオンライン百科事典であり、2010 年 8 月時点で、英語版で 337 万件、日本語版で約 69 万件の記事が登録されている。これらの記事は、様々なユーザが加筆や修正を行いながら、作成されている。変更された記事は記録され、誰でもすべての変更履歴を閲覧することができる。Wikipedia では、コンテンツのクロールを禁じているが、その代替手段として、各記事の内容を含め変更履歴の一括ダウンロードが可能である。変更履歴は、XML 形式で記述されており、記事ごとの変更内容と時間、編集したユーザ名を取得することができる。

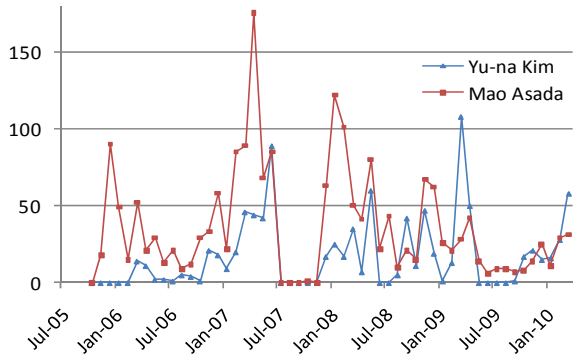

図1. 更新履歴の推移

本研究では、Wikipedia のタイトルごとに、変更頻度を一定期間で集計し、その合計をタイトルの語の注目されている尺度（注目度）として利用する。例えば、フィギュアスケートの浅田真央選手と金妍兒（キムヨナ）選手の変更履歴の推移（月単位）は図 1 のようになる。大きな試合がある 3 月、6 月、12 月など注目されている時期には、変更の頻度が上昇する傾向がわかる。このように更新履歴から得られる注目度を、番組アーカイブの情報のみから求めた検索結果のリランキングに用いる。

2.2 更新履歴のノイズ対応

同一ユーザの頻繁な編集

変更履歴には、同一 ID による短期間の集中的な書き込みや、ユーザ間での編集合戦(Edit War)など、本来、注目度としては複数回カウントすべきでないログが含まれている。そのため、一定期間に集中する同一ユーザによる書き込みを更新数から省く。今回は、1 日毎にユニークな編集ユーザ数をカウントすることにより、ノイズの除去を行う。

保護期間の補完

編集合戦や不正な更新が続くと、管理者により誰も編集ができなくなる保護期間が設けられる場合がある。図 1 の 07 年 06 月から 11 月までが両記事ともに保護期間にあたる。保護期間中の変更の頻度は 0 となるが、本来であれば多くの変更が行われる可能性がある。そこで、保護期間の直前と直後の平均を頻度とする補完措置を行う。

† NHK 放送技術研究所, NHK Science and Technology Research Laboratories
‡ 国立情報学研究所 National Institute of Informatics

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.

3. 注目度を利用した番組検索結果のリランキング

3.1 番組概要を利用した番組検索

我々は過去に番組概要の類似性を利用した関連番組検索手法を提案した[2]。手法では Okapi BM25 [8]に基づいた類似度計算に固有表現の種類による重みを導入し、関連番組の検索を行う。番組データベースの番組概要に出現する n-gram を t_i、その総数を I とすると、クエリ番組 Q とデータベースのある番組 D の n-gram ベクトル間の関連性のスコアは式(1)で求められる。Q, D の各要素 $q(t_i)$, $d(t_i)$は BM25 に基づき式(2)(3)のように定める。$W_{ne}(t_i)$は固有表現の種類（人物名、組織名、地名など）により決定する。

$$S = \sum_{i=1}^{I} W_{ne}(t_i) \cdot q(t_i) \cdot d(t_i) \quad (1)$$

$$q(t_i) = \frac{(k_3+1)tf_q(t_i)}{k_3 + tf_q(t_i)} \quad (2)$$

$$d(t_i) = \frac{(k_1+1)tf_d(t_i)}{k_1((1-b)+b\cdot\varepsilon/\eta)+tf_d(t_i)} \log\left(\frac{M - m(t_i)+0.5}{m(t_i)+0.5}\right) \quad (3)$$

M は番組アーカイブの番組数、$m(t_i)$ は t_i を含む番組数を示す。ε, η はある番組概要の長さ、データベースに含まれる番組概要の平均の長さを示す。k_1, k_3, b は調整用のパラメータである。

3.2 更新履歴に基づく注目度の利用

提案手法では、前節で説明した手法の固有表現による重み $W_{ne}(t_i)$の代わりに、更新頻度に基づく注目度 $A(t_i)$を重みとして利用し、注目度によるリランキングを行う。

$$S = \sum_{i=1}^{I} A(t_i) \cdot q(t_i) \cdot d(t_i) \quad (4)$$

注目度 $A(t_i)$は、各記事の更新頻度 $freq(t_i)$をそのまま注目度とすると、更新頻度の幅が大きいため、式(5)を用いる。

$$A(t_i) = \log_r (r + freq(t_i)) \quad (5)$$

Wikipedia の記事の更新では、人物名や組織名などのカテゴリの種類により、その傾向が異なる。そのため、更新頻度をどの程度を考慮するかを定めるパラメータとして底 r を用いる。

4. 動作実験

提案手法の効果を確認するため、番組アーカイブの 9471 番組を対象に、注目度を利用した提案手法と、3.1 節で述べた既存手法との比較を行った。クエリとした番組は、ドラマ（現代、時代劇）、スポーツ、報道番組の４種類を選び、それぞれ 20 位までの関連番組を検索した。注目度に利用する表現は人物名を対象とした。式(5)の底 r は２とし、最新の２年の変更履歴を注目度として利用した。また、ベースライン手法として、Okapi BM25（固有表現と注目度の重み付与なし）と tf-idf 重みによるコサイン類似度を用いた。

評価手法には、DCG (Discounted cumulative gain) [9]を利用した（４段階評価０-３）。実験の結果（表１）、提案手法によるリランキングの結果が、すべての順位の DCG で既存手法及びベースライン手法を上回った。これは、ユーザが注目している番組が上位にランキングされたためと考えられる。例えば、図２のクエリ「篤姫」（下線は人物名）での結果では、「龍馬伝」や「そのとき歴史が動いた」などの番組が提案手法によるリランキングで上位となった。Wikipedia の更新履歴から得た「坂本龍馬」等の注目度により、これらの番組の順位が上昇したためと考えられる。

将軍・家茂の元に、上洛と攘夷実行を求めて京から勅使が訪れます。・・・・和宮は家茂の身を案じて上洛に反対し、後押ししたのが天璋院だと知って強い敵対心を抱きます。勝麟太郎を斬るためにやってきた坂本龍馬は、勝の進歩的な考えに感銘を受け、弟子になりたいと志願します。

図 2. クエリとした番組概要例

	Proposed	BM25+NE	BM25	tf-idf
DCG_3	9.32	8.46	8.58	8.50
DCG_5	12.93	11.58	11.72	11.48
DCG_{10}	18.04	16.60	16.53	15.41
DCG_{20}	24.05	22.34	22.32	21.57

表1. 各順位での評価

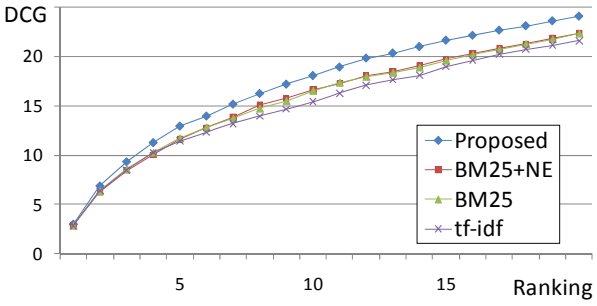

図 3. 実験結果

5. まとめ

更新履歴に基づく注目度による関連番組検索のリランキング手法を提案した。評価実験の結果、Wikipedia の更新履歴を利用した提案手法が、番組アーカイブの情報のみから求めた検索結果を改善する可能性を示した。今後、注目度に利用する変更履歴の期間を変更し、その動作の変化を検証する予定である。

参考文献

[1] https://www.nhk-ondemand.jp
[2] J. Goto, H. Sumiyoshi, M. Miyazaki, H. Tanaka, M. Shibata and A. Aizawa: Relevant TV program retrieval using broadcast summaries, In Proc. of IUI2010, pp.411-412, (2010).
[3] D. Beeferman and A. Berger: Agglomerative Clustering of Search Engine Query Log, Proc. of SIGKDD 2000, pp.407-416, (2000)
[4] M. Eirinaki and M. Vazirgiannis: Web Mining for Web Personalization, ACM Transactions on Internet Technology, Vol.3, No.1, pp.1–27, (2003)
[5] K. Ali, W. Stam: Making Show Recommendations Using a Distributed Collaborative Filtering Architecture. In Proc. of the SIGKDD 2004, pp.394-401, (2004).
[6] E. Yamangil and R. Nelken: Mining Wikipedia Revision Histories for Improving Sentence Compression. Proc. of ACL08-HLT, pp.137-140, (2008)
[7] H. Zeng, M. Alhossaini, L. Ding, R. Fikes1 and D. L. McGuinness: Computing trust from revision history. Intl. Conf. on Privacy, Security and Trust, (2006).
[8] S. Robertson and S. Walker: Okapi/Keenbow at TREC-8, Proc. of the 8th Text Retrieval Conference, (1999).
[9] K. Jarvelin and J. Kekäläinen: IR Evaluation Methods for Retrieving Highly Relevant Documents, Proc. of SIGIR 2000, pp.41-48, (2000)

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.