2F1-04

ユーザの着眼点を考慮したスニペット生成

Snippet generation reflecting diverse aspects

田中 陽子[*1] 廣嶋 伸章[*1] 別所 克人[*1] 小池 義昌[*1] 片岡 良治[*1]
Yoko Tanaka Nobuaki Hiroshima Katsuji Bessho Yoshimasa Koike Ryoji Kataoka

[*1]日本電信電話株式会社　NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

Snippets are one of the most important elements of web search engines because they help users decide the relavancy of the returned documents. Conventional methods generate snippets based on the relevancy to the queries and the fidelity to the original documents. Hence, they provide the same snippets regardless of the user. However, as users may have different aspects of interest, it is desirable to customize snippets for each user. We propose a method to generate snippets that reflects users' diverse aspects by incorporating a preference model which represents users' aspects. The results of evaluation show reasonable efficacy of our method.

1. はじめに

近年，インターネットの普及に伴い，様々な情報を持つ膨大な文書を誰でも簡単に手に入れることができる環境が整ってきており，その中から個々のユーザが求めている情報を効率的に探す試みが行われている．求めている情報を探す最も主流な手段として，クエリを入力して関連する文書を得る検索エンジンの利用がある．ユーザが検索によって情報を得るために，システムは 1) 検索対象の文書群からクエリを含む文書を選択し，2) クエリについて記述された文書の概要文 (スニペット) を生成して提示することが必要である．1) については検索ランキングの問題として様々な研究が行われてきた．一方，2) については文書の query-biased な要約を生成する問題である．この問題に関する研究として，文書中の単語の組み合わせを要約候補とし，要約候補に含まれる単語の文書中での出現確率と，クエリ単語の要約候補中での出現確率から最適な要約を生成する手法がある ([A.Berger 00])．この従来技術を利用した場合，従来の検索エンジンのように，同じクエリを入力したユーザに対しては同じ要約が生成され，スニペットとして提示される．

しかし，同じクエリを入力しても，ユーザによって着目する部分（以下，着眼点と呼ぶ）が異なる場合がある．例えば,「有馬　温泉」という同じクエリでも，ある人は値段に関する情報に着目しているが，ある人は雰囲気に関する情報に着目している．このように，年齢や家族構成といったユーザの属性や，安い商品を買いたいといった趣向等によって，注目する部分や意思決定の際に重要視する部分が異なる．従来の要約には，着眼点に関する情報が必ずしも含まれていないため，ユーザは検索結果の文書群の中から自分の着眼点について述べられている文書を探す作業を複数回繰り返さなければならなかった．そこで，本稿ではユーザの個々の着眼点を考慮した，Query-biased な要約の生成手法を提案する．着眼点による選好性モデルを導入することで，クエリが同じでも，個々のユーザが求めている情報をより多く取り入れた要約の生成を目指す．

2. 提案手法

この章では，ユーザの着眼点を考慮した query-biased な要約の生成手法について述べる．

連絡先: tanaka.yoko@lab.ntt.co.jp

2.1 概要

本研究では，与えられた要約の文字数制限を満たす文書内の文の組み合わせを要約の候補とし，複数生成された要約候補の中から最適なものを選択する．着眼点を考慮した query-biased な要約とは，文書 d に対してクエリ q，着眼点 a が与えられた時の最適な要約 s である．これを式にすると，

$$s^* = \arg\max_s P(s \mid d, q, a)$$

と表される．これにベイズの定理を適用すると，

$$P(s \mid d, q, a) = \frac{P(s \mid d)P(a \mid s, d)P(q \mid s, d, a)}{P(a \mid d)P(q \mid d, a)}$$

となる．ここで，$P(a \mid s, d)$ において d は a を決める際に影響を及ぼさないと考えられる．同様に，$P(q \mid s, d, a)$ において d, a は q を決める際に影響を及ぼさない．また，分母は s に依存しないことから，

$$s^* = \arg\max_s P(s \mid d)P(q \mid s)P(a \mid s)$$

と表すことができる．この 3 つの確率の積が要約候補の中で最も大きいものが最適な要約となる．

2.2 文書に対する忠実性モデル

要約は文書の代用として利用されるものであり，文書の概要を正しく表した内容である必要がある．$P(s \mid d)$ は，文書に対して，生成した要約候補の単語列が元の文書の内容をどれだけ忠実に再現しているかを表していると考えることができる．要約候補の単語列 s に含まれる名詞を $s_1, s_2, ..., s_m$ とし，それぞれ要約中に $c_1, c_2, ..., c_m$ 回現れるとする．また，文書中の全名詞数を n とすると，名詞 s_i が文書中に k_i 回出現する場合の確率は $p_i = k_i/n$ と表される．要約中の全名詞数は c 個とする．忠実性モデルは，各名詞 s_i が確率 p_i で c_i 回選ばれた場合の多項分布に従うものとしてモデル化できる．

$$P(s \mid d) = \frac{c!}{\prod_i^m c_i!} \prod_i^m p_i^{c_i}$$

しかし，文書中に長さが不均一な文が含まれている場合，この方法では長さが短く，名詞の数が少ない要約候補の確率が高く

なる傾向がある．そこで，ゼロ頻度問題に用いられるラプラス法によって，文の長さに対する確率の補正を行う．名詞数が最も多い単語列と着目している単語列の名詞数の差を D とすると，D 種類の仮想の名詞がそれぞれ文書中に 1 回ずつ出現していると考える．文書中の名詞の種類が V 個であるとすると，名詞 s_i が文書中に k_i 回出現する場合の確率を，

$$p_i' = \frac{k_i + 1}{n + V + D}$$

とする．仮想の名詞については $k_i = 0$ とする．この確率で計算することで，単語列の長さによる忠実性の差が補正される．

2.3 クエリとの関連性モデル

文書中に，与えられたクエリと関連を持つ部分と持たない部分がある場合，生成された要約がクエリと関連を持つ部分をどれだけ含んでいるかが重要である．$P(q \mid s)$ は要約がどれだけクエリと関連があるかを表したものと考えることができる．クエリの全名詞数を c 個とし，m 種類の単語からなるクエリ $q_1, q_2, ..., q_m$ について，q_i が n 個の単語からなる要約候補の単語列の中に確率 $p_i = k_i/n$ で c_i 回出現する場合の多項分布に従うものとしてモデル化できる．

$$P(q \mid s) = \frac{c!}{\prod_i^m c_i!} \prod_i^m p_i^{c_i}$$

2.4 着眼点による選好性モデル

従来技術では，前述のような文書への忠実性とクエリとの関連性のみを元に要約を生成していた．しかしこれでは，ユーザの着眼点が反映されず，ユーザが求めている情報が要約に取り入れられない可能性が高い．そこで，我々はユーザの着眼点をクエリと同様に入力として与えた際，生成された要約がユーザの着眼点をどれだけ表現しているか重要であると考え，$P(a \mid s)$ は，これを表したものと考えた．着眼点を，物事を判断する際に注目する点や重要視する点を 1 つ以上の単語で表現したものとする．一般に，着眼点は文書中に単語がそのまま含まれているものばかりではないため，着眼点の単語の出現確率ではなく各着眼点における類似度の平均に基づく多項分布に従うものとする．要約候補の単語列に含まれる名詞と着眼点の単語について，単語間類似性判定データベースである概念ベースを参照して単語間の類似度を求める ([別所ら 06])．概念ベースとは，コーパスにおける単語同士の共起頻度を記録した共起行列に対し，特異値分解を行い，単語を次元数の縮退したベクトルで表現した概念ベクトルのデータベースである．着眼点 $a_1, a_2, ..., a_n$ と要約に含まれる名詞 $s_1, s_2, ..., s_m$ について，それぞれの概念ベクトルから全ての着眼点 a_i と名詞 s_j の単語間の類似度を求める．ここで，単語間の類似度は単語の概念ベクトルのコサイン距離とする．着眼点 a_i について値が大きい順から v 個の類似度を $t_{i1}, t_{i2}, ..., t_{iv}$ とすると，選好性モデルは，これらの平均の全着眼点における積に従うものとしてモデル化できる．

$$P(a \mid s) = n! \prod_i^n \frac{1}{v} \sum_l^v t_{il}$$

3. 評価

提案手法の有効性を調べるために，忠実性と関連性のみを用いる従来手法との比較を行った．要約の対象とする文書は，人によって着眼点が異なるであろうと考えられる，「旅行」と「食べ物」に関する文書とした．まず，旅行と食べ物それぞれに関するクエリ（例：「有馬　温泉」「惣菜　取り寄せ」等）を 3 種類ずつ，計 6 種類用意し，商用検索エンジンを使ってそれぞれのクエリでウェブ検索した結果の上位 25 件，合計 75 件を要約対象の文書とした．次に，3 人の被験者に対し，値段，雰囲気等，着眼点になりうると考えられる 6 項目について，旅行と食べ物でそれぞれ意思決定をする際に重要視するものを順位付けをしてもらい，被験者それぞれの上位 3 項目に関する単語（例：「雰囲気，値段，家族」等）を提案手法で用いる着眼点とした．選好性モデルにおいては，$v = 2$ とした．75 件の要約対象の文書に対し，着眼点を用いた提案手法による要約と，着眼点を用いない従来手法による要約をそれぞれ作成し，被験者にどちらの要約が自分にとってふさわしい要約であるかアンケートをとった．結果を表 1 に示した．着眼点 A_1 から A_6 についての提案手法と従来手法による要約を比較した際，それぞれの手法で生成された要約が異なる場合と全く同じ場合がある．異なる要約が生成されたもののうち，提案手法のほうがふさわしいとされたものを○，従来手法のほうがふさわしいとされたものを×，優劣がつけ難いものを△とした．この結果から，着眼点によっては従来手法による要約がわずかに上回ったものもあったが，ほとんどの着眼点において，提案手法によってよりユーザにふさわしい要約を生成することができることが分かった．一方で，多くの要約が従来手法と提案手法で差が見られなかった．優劣がつけ難いと判断された要約の中にも，主要な文は同じで，5 文字程度の短い文のみが異なり，要約の内容に差異がないものが多かった．実際に被験者が，元の文書から文を選択して 100 文字程度の要約を作った結果，着眼点に関する情報を含むにも関わらず，提案手法には含まれなかった文を選択していることが分かった．今後は，被験者が作った要約を参考に，選好性モデルの改善のための方法を検討する．

表 1: 評価結果

クエリ		旅行			食べ物		
着眼点		A_1	A_2	A_3	A_4	A_5	A_6
異なる要約	○	**8**	**10**	**8**	**10**	7	**9**
	×	0	2	2	6	**8**	4
	△	12	13	14	7	5	13
同じ要約		55	50	51	52	55	49

4. おわりに

本研究では，ある文書に対してクエリだけでなくユーザの着眼点を考慮した要約の生成手法を提案した．従来手法と比較した結果，提案手法により，ユーザが重要視する情報を取り入れ，ユーザそれぞれにふさわしい要約の生成が可能であることが示唆された．今後は，実際に人手で作成した要約に近づけるように要約の質を向上する方法を検討する．

参考文献

[A.Berger 00] A.Berger, V.O.Mittal: Query-relevant summarization using FAQs(2000), ACL '00 Proceeding of the 38th Annual Meeting on Association for Computational Linguistics.

[別所ら 06] 別所克人，古瀬蔵，片岡良治: 単語と意味属性の共起に基づく概念ベクトル生成法 (2006)，人工知能学会全国大会 2006.