情報検索のためのクエリに基づく文書自動要約

桜井 俊彦　内海 彰
電気通信大学大学院 電気通信学研究科 電子情報学専攻
toshi@utm.se.uec.ac.jp ，utsumi@se.uec.ac.jp

1　はじめに

情報検索において検索結果を表示する際，ほとんどのシステムではユーザが入力したクエリ (ユーザの検索要求) が含まれている箇所の抜粋を提示している．ユーザはこのわずかな情報を頼りに，類似した文書を多く含む検索結果の中から求める情報を含む文書を探すことになる．しかし単純にクエリを含む箇所の抜粋だけからでは，どの文書ががユーザにとって有用な情報かどうかを判断することは困難である．そこで，代わりにクエリに基づく要約を提示することでユーザの検索の負担を軽減できると考えられる．

情報検索のための要約作成手法として，単純にクエリを含む文を重要文として計算する手法 [1] が提案されているが，実際にこの手法で生成された要約はクエリを含む箇所の抜粋とあまり変わらない．また，検索された文書間の表層的類似性を適切に説明する語を重要語とする手法 [2] も提案されているが，クエリを検索結果を得るためにしか用いておらず，検索の質と結果の文書数に大きく依存しているため，検索システムとの統合が必要であるという問題がある．

本研究は，ユーザの求める情報かつ文書に特徴的な情報を含む要約作成を目指す．これを実現するために二段階の要約作成手法を提案する．まずユーザの求める情報を抽出するため，クエリと共起する語やクエリと意味的な関連度が高い語に着目してクエリに関する要約を作成する．次に検索結果の中で他の文書との違いをユーザに提示するため，クエリとの関連が低い語などに着目して原文に特徴的な内容に基づく要約を作成する．以上の二つの要約を合わせて概要把握のためのクエリに基づく要約とする．

2　文書自動要約

2.1　要約の種類

要約にはその目的に応じて二種類の要約がある [3]. 一つは，原文を参照するかどうかを判断するために用いられる indicative な要約であり，もう一つは原文の大意を含み，原文の代わりとして用いられる informative な要約である．本研究では検索結果の中から求める情報を含む文書を探すための indicative な要約作成を目指す．

2.2　要約手法

一般的に人間が要約を作成する過程は大きく以下の三つのステップに分けられる [4].

1. 文書の解釈
2. 解釈に基づく要約の内部表現への変形
3. 重要部分からの要約文生成

しかし，現在の技術では文書の内容や意味をコンピュータに理解させるには莫大な知識が必要となり困難である．またコンピュータが自動的に意味の通る文を生成することは現状では不可能である．よって自動要約研究では上記の解釈から変形までの過程を文書から得られる表層的な情報 (語の出現頻度，段落等の位置など) を基に重要箇所を抽出することで実現している．また，要約文生成に関しては文を一つの単位とし，先に述べた情報を基に文単位で重要度を付与し，重要な文を選択しそれらを抜粋することで要約を作成するのが一般的かつ有効な手法である．本研究ではこの重要文抽出型の手法に従って要約を作成する．

3　クエリに基づく要約作成手法

3.1　手法概要

検索結果はクエリに関する類似した文書の集合である．ユーザはクエリを含む箇所ばかりを提示されても，どの文書が求める情報を含んでいるかを的確に判断することは困難な作業となる．そこで本研究ではクエリを直接用いることを避け，分類語彙表や文書集合を用いて語の重要度計算を行い，クエリとの関連が高い情報を抽出する．さらに，文書集合中の他の文書との違いを要約に反映させるため，原文に特徴的な情報を抽出する．

以下の過程を経て文書 D の要約 $Summary(D)$ を作成する．手法の概念図を図 1 に示す．

Step 1 文書 D のクエリに関する要約 $Summary_Q(D)$ を作成する．

Step 2 Step 1 で要約として選択されなかった文書 D' から文書 D に特徴的な内容に基づく要約 $Summary_C(D)$ を作成する．

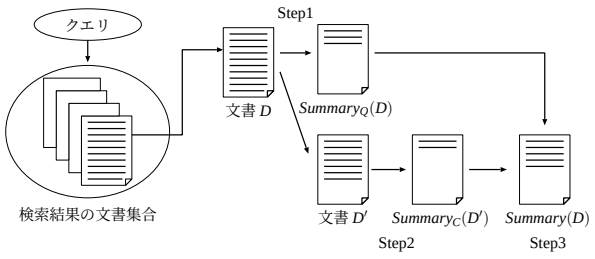

図 1: 概要把握のための自動要約概念図

Step 3 $Summary_Q(D)$ と $Summary_C(D)$ として選択された文を原文 D で出現する順番に並べて文書 D の要約 $Summary(D)$ とする．

Step 1 では単純にクエリを含む文を重要とみなすのではなく，クエリの関係する情報を積極的に利用して，クエリに関する要約 (*Query-based Summary*) を作成する．次に Step 2 では文書集合中の他の文書には含まれない文書 D に特有の情報を提示するために，文書 D に特徴的な内容に基づく要約 (*Content-based Summary*) を作成する．そして Step 3 で要約 ($Summary(D)$) として出力する．以下の節で Step 1 と Step 2 の詳細を示す．

3.2 クエリに関する要約

以下に示すアルゴリズムによって，クエリに関する要約を作成する．なおこの手法の概念図を図 2 に示す．

Step 1-1 文書 D における語 t の重要度 $W_D(t)$ を以下の三つの情報を用いて計算する．

- 語 t の出現頻度 $TF_D(t)$
- 語 t のクエリとの意味的な関連度 $SQN_D(t)$
- 語 t のクエリとの共起度 $COR_D(t)$

なお，ここで言う語とは，形態素解析ソフト「茶筌」で名詞 (' 非自立', ' 特殊', ' 副詞可能', ' 助動詞語幹' は除く) と出力される語である．

Step 1-2 語の重要度を用いて文 S の重要度 $I_D(S)$ を計算する．

Step 1-3 与えられた要約の長さ (本研究では 150 文字以内) を満たすまで重要度の高い順に文を抜粋し，クエリに関する要約 $Summary_Q(D)$ とする．

3.2.1 語の重要度の計算方法

語の出現頻度

文書中で頻出する語は文書を特徴付ける語であるとされる．出現頻度による重要度 $TF_D(t)$ は文書 D 中で語 t が出現する回数とする．

クエリとの意味的な関連度

分類語彙表を用いて類似度に応じて得点を決定する．分類語彙表 [5] とは国立国語研究所が研究目的に提供している，単語を意味的に分類した辞書である．語 t とクエリの分類番号の上位 4 桁が同じときには，どの分類まで一致しているかに応じて表 1 のように $SQN_D(t)$ を決定する．クエリが複数ある場合はそれぞれのクエリに対して独立に計算される．

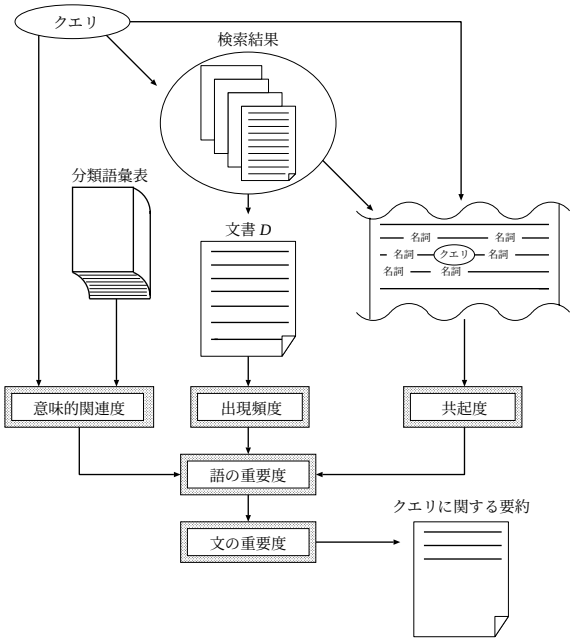

図 2: クエリに関する要約作成概念図

クエリと共起する語

対象となる文書中でクエリと共起する語が文書集合中でも同じようにクエリと共起していた場合，共起している語はクエリを間接的に特徴付ける語であると考えられる．そこで文書 D においてクエリと共起する語 t が全文書中でクエリと共起する回数 $COF(t)$ を求め，これを基に式 (1) に従って重要度 $COR_D(t)$ を計算する．なお本研究ではクエリの前後 10 語以内に出現する語を共起語とする．

$$COR_D(t) = log_2\left(\frac{COF(t)}{TF_D(t)} + 1\right) \quad (1)$$

式 (1) では文中での語 t の出現頻度が高ければクエリと共起する回数も増えることが予想されるため，$TF_D(t)$ で正規化している．ここで一つの文書内に同じクエリが

表 1: クエリに関連する語の重要度計算

一致している桁数	具体例	$SQN_D(t)$
分類番号上位 4 桁	貿易 1.4760,2,1,3 売買 1.4761,1,1,2	2
分類番号	サッカー 1.3374,8,2,1 テニス 1.3374,7,7,1	4
分類番号 段落番号	野球 1.3374.9,1,2 奪三振 1.3374,9,14,1	7
分類番号 段落番号 段落内番号	議会 1.2730,1,1,2 国会 1.2730,1,1,3	10

2 つ以上存在する場合やクエリが複数ある場合，共起範囲が重なる可能性がある．これらの場合すべて重複して数え，$COF(t)$ を求める．

以上の結果に基づき文書 D における語 t の重要度 $W_D(t)$ を次式で計算する．

$$W_D(t) = TF_D(t) + SQN_D(t) + COR_D(t) \quad (2)$$

3.2.2 文の重要度の計算方法

文書 D における文 S の重要度 $I_D(S)$ を次式で計算する．式 (3) では文の長さによって差が出ることを防ぐために，文 S に含まれる語数 n で正規化している．

$$I_D(S) = \sum_{t \in S} \frac{W_D(t)}{n} \quad (3)$$

3.3 原文に特徴的な内容に基づく要約

3.1 節で述べた通り，文書集合中の他の文書との違いを要約に反映させるために原文中でクエリとの関連が低い部分に着目して要約を行う．ここでは $Summary_Q(D)$ に含まれていない語を重要語とみなし，重要文抽出型の手法で要約を作成する．

原文から $Summary_Q(D)$ を除いた文書を D' とする．アルゴリズムは以下の通りである．

Step 2-1 文書 D' における語 t の重要度 $W_{D'}(t)$ を以下の二つの情報を用いて計算する．

- 文書 D を特徴付ける語の重要度 $PTF_{D'}(t)$
- 文書集合における語の文書出現頻度 $IDF_{D'}(t)$

Step 2-2 式 (3) によって文の重要度を計算し，重要度の高い文を $Summary_C(D)$ とする (要約の長さは 75 文字以内).

3.3.1 語の重要度計算方法

文書 D を特徴付ける語の重要度

文書 D' に特有の情報を決定するため，クエリに関する要約 $Summary_Q(D)$ に含まれない語に着目する．語 t の D'，$Summary_Q(D)$ での出現頻度を含まれる語数で正規化した値をそれぞれ $TF_{D'}(t)$，$TF_{S(Q)}(t)$ とし，次式により文書 D' における特徴的な語の重要度 $PTF_{D'}$ を計算する．計算例を図 3 に示す．

$$PTF_{D'}(t) = \begin{cases} TF_{D'}(t) - TF_{S_Q(D)}(t) & TF_{D'}(t) > TF_{S_Q(D)} \\ 0 & TF_{D'}(t) \leqq TF_{S_Q(D)} \end{cases} \quad (4)$$

文書集合における語の文書出現頻度

文書集合中において，ある特定の文書にのみ出現する語は文書集合中における特徴的な語とみなすことができる．文書集合の文書数を N，文書集合において語 t を含む文書数を $df(t)$ とすると $IDF_{D'}(t)$ は次式で計算される．

$$IDF_{D'}(t) = \log \frac{N}{df(t)} + 1 \quad (5)$$

以上の結果に基づき文書 D' における語の重要度を次式で計算する．

$$W_{D'}(t) = PTF_{D'}(t) \times IDF_{D'}(t) \quad (6)$$

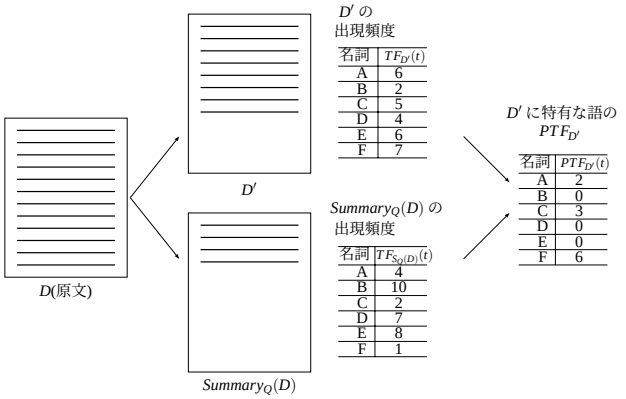

図 3: 文書 D を特徴付ける語の重要度計算例

4 評価

本研究で提案した手法の有用性を検証するためにアンケートに基づく評価を行う．第 4 回 NTCIR ワークショップにおけるテキスト自動要約タスク (TSC-3) において配布されたテストコレクションを使用する．データはクエリと新聞記事 (毎日新聞 1998-1999，読売新聞 1998-1999) から検索された文書集合のセットで構成される．今回は 4 つのセットを選択し，さらにそれぞれの文書集合から 4 つの文書を選択してアンケートに用いるデータとする．提案手法，比較手法によって作成される要約をそれぞれ X_D，Y_D とする．まず被験者に検索結果として，4 つの文書の要約 (X_D or Y_D) を提示する．質問は，要約に対して原文はクエリを満たす内容を含んでいるか (Q1) を 4 段階 (1：含んでいると思わない〜4：含んでいると思う) で評定してもらう．次に原文を提示した上で，要約が原文を見るかどうかを判断する材料としてふさわしいかどうか (Q2) を同じく 4 段階で評定しても

表 2: アンケートによる評価結果

	Q1		Q2	
クエリ	本研究	比較手法	本研究	比較手法
A	2.95	2.45	3.25	2.70
B	2.95	2.75	3.20	3.00
C	3.00	2.45	3.45	2.70
D	3.35	1.40	3.20	2.20

らう．全部で 4 つのクエリを用意し，それぞれの要約に対し 5 人ずつに評定してもらう．比較手法はクエリを直接重要度計算に用いる手法 [1] を参考に，TF-IDF 法で重要度が付与された文に対してクエリを含んでいたら重要度の値を 3 倍にするという計算方法で要約を作成する．表 2 は各クエリに対する評価値の平均を示す．

Q1，Q2 ともに全てのクエリに対して，提案手法が比較手法より平均値が高くなっていることがわかる．クエリ B では選択された原文にユーザが求める情報が頻出していることが，平均の差が小さくなった原因だと考えられる．原文を参照してから評価した結果，クエリ D 以外は評価値が上がっている．これは被験者が原文に求める情報が存在しないと判断した際に，要約がふさわしいと評価したためである．これは，質問が具体的すぎて要約に答えが含まれているかどうかが大きな評価基準になってしまったためだと考えられる．

この実験より，情報検索の結果表示部分の要約として，単純にクエリを重み付けする従来の手法より本研究で提案する手法の方が適していることがわかる．

5 考察

重要語の分布

要約結果にクエリが含まれる頻度を調査した．4 章と同じテストコレクション (クエリと文書集合の 30 セット) を対象に，各文書を提案手法で要約した結果 $Summary_Q(D)$, $Summary_C(D')$ それぞれに含まれる重要語，クエリの数の 30 セット (全文書数 352) の平均を表 3 に示す．ここで重要語は $W_D(t)$ が文書 D における $W_D(t)$ の総和の 2%以上を占める語とする．$Summary_Q(D)$ はクエリと重要語 (クエリに関連する語) が頻出している．これはクエリに関する情報を多く含むことを意味する．逆に $Summary_C(D')$ ではクエリも重要語も頻度が低くなっている．この二つの要約を合わせることで，ユーザにクエリに関する情報を提示し，かつ他の文書との違いを提示することを実現しているといえる．図 4 は「パプアニューギニア 地震 津波 被害」というクエリを入力した場合の出力例である．破線部はクエリに関する要約 $Summary_Q(D)$ として選択された文で，クエリ自身を多く含んでいる他，重要語である「災害」,「調査」等の語を多く含んでいることがわかる．原文に特徴的な内容に基づく要約 $Summary_C D'$ として選択された文にはクエリを含まず，表層的にはクエリと関連が低いと思われる箇所が抽出されていることがわかる．

表 3: 重要語の分布

	$Summary_Q(D)$	$Summary_C(D')$
クエリ	3.69	0.76
重要語	11.50	2.98

パプアニューギニア北西部に大きな被害をもたらした津波災害の実態を解明するため，国内外の研究者らが調査団を結成，３０日から現地入りする．米国，オーストラリア，ニュージーランドの７人も加わる．８月９日まで現地に滞在し，直撃を受けた西セピク州沿岸一帯で津波の高さや時間帯，音の有無などの聞き取り調査を実施．地形や地震記録などを基に津波の規模や大災害となった原因を調べる．

図 4: 要約出力例

Web への実装

本研究では WWW 検索を想定しているため，CGI を用いて Web 上に実装した．しかし，Web 上の HTML 文書では表や図といった有用な情報を含んでいるが文になっていない箇所が多く存在する．このような形式の整っていない文をどのように処理するかということは非常に困難な課題である．解決方法としては，HTML タグ情報を用いて文に近い単位を構築することや，言い換えなどの技術を適用することが挙げられる．

6 おわりに

本研究は情報検索の検索結果表示部分のための要約生成を目指し，概要把握という観点からクエリに関連する情報を抽出し，他の文書との違いを要約に反映させるために原文に特徴的な情報を抽出するという手法を提案した．評価の結果，既存の手法と比較して提案手法が有用であることがわかった．今後は Web への実装と精度の向上が課題である．

また本研究での重要度計算を応用して，複数文書要約や文書の中から知識を発掘するテキストマイニングの研究に発展させていきたいと考えている．

参考文献

[1] Tombros,A.and Sanderson,M.: Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, 2-10. (1998).

[2] 森 辰則: 検索結果表示向け文書要約における情報利得比に基づく語の重要度計算，自然言語処理，Vol.9, No.4，pp.3–32 (2002).

[3] 奥村 学，難波 英嗣: テキスト自動要約に関する研究動向 (巻頭言に代えて)，自然言語処理, Vol.6, No.6, pp.1–26 (1999).

[4] Mani,I.: Automatic Summarization,John Benjamins Publishing Company, (2001).

[5] 国立国語研究所:「分類語彙表」形式による語彙分類表 (増補版), (1996).