情報利得比に基づく重要語抽出による情報ナビゲーション

吉田和史 † 塩田好伸 ‡ 森辰則 ††
† 横浜国立大学 大学院 工学研究科 ‡ 横浜国立大学 工学部 †† 横浜国立大学 大学院 環境情報研究院
E-mail: {kazu,yoshinobu,mori}@forest.eis.ynu.ac.jp

1 はじめに

近年，WWW 検索エンジンのような情報検索システムが広く利用されるようになり，検索要求に関連のある文書を容易に得る事が出来るようになった．しかし，検索要求に関連性の低い文書を完全に排除できない，検索結果文書の構造化がなされていないなどの問題点があり，検索結果文書から利用者が真に必要とする文書を効率よく入手するのは困難である．

有効な方策としては，検索された各文書の要約の提示することによる元文書参照時間の削減，検索結果文書をクラスタリングすることによる関連文書と不要文書の分類などが従来提案されている．これらの方策に共通する必要不可欠な技術は検索結果文書集合を考慮した重要語の抽出である．そのような重要語抽出手法の代表は検索要求中の語の重要度を高くする手法であり，これを自動要約に利用したものが Query-biased Summarization[1] である．この手法は直観的ではあるが，検索エンジンによる各種フィードバック等の工夫が反映されないという問題点がある．

そこで，我々は検索質問ではなく，検索文書間の関係から重要語を抽出することを検討している．これは，複数文書間の類似性構造というマクロな情報を，語の重要度というミクロな情報に写像する新しい手法である．この手法では，文書分類の過程で得られた文書間の類似性構造を適切に説明する度合を語の重要度とする．

本稿では，この重要語抽出手法が方法論として文書分類と自然に融合している点に着目し，文書分類に基づく情報ナビゲーションと文書要約を同時に兼ね備えた情報ナビゲーションシステムを提案する．特に，情報ナビゲーションに必要な説明記述の生成について各種手法を比較検討する．

2 情報検索結果に対するナビゲーションにおける課題

必要な文書を検索文書の中で絞り込む際に重要な事柄は，検索要求と関連がある文書と無い文書の区別，ならびに，関連がある文書内での更なる詳細な分類である．そこで，文書間の類似度に基づいた文書分類を基礎とする情報ナビゲーションを考える．分類提示する手法としては，次節で述べる Scatter/Gather 流の動的なクラスタリングに基づく対話型インターフェース [2] が非常に柔軟性に富むのでこれを採用するが，そこで問題になるのが，分類された構造 (クラスタ) を利用者に対してどのように説明するかという点である．

各クラスタに対する説明記述を生成する手法としては，クラスタ内の重要語 (キーワード)，クラスタ内の文書集合を代表する文書のタイトル，あるいは，文書の要約を提示する等，様々なものが考えられる．本稿では，クラスタ内の文書数が大きい段階ではクラスタ全体のキーワードを提示し，十分に絞りこまれた段階で文書のタイトルとともに要約を提示する方式を次節以降で検討する．

3 提案手法

先に述べた通り，Scatter/Gather 流の対話型インターフェースを利用する．Scatter/Gather 手法では，まず，検索結果文書を決められた数のクラスタに分類する．クラスタリングは，TFIDF 法に基づくベクトル空間法を距離尺度として，グループ平均法により行われている．次に，分類された各クラスタに対し，クラスタ内の文書を説明するキーワードを付与する．利用者はその説明記述を基に，必要と思われる文書集合を選択する．システムは，選択されたクラスタを併合し再びクラスタリングを行なう．そして，以上の処理を繰り返す事により文書集合を絞り込んでいく．

さて，Cutting ら [2] においては，Scatter/Gather のアルゴリズムの提案に焦点があり，利用者インタフェースとして重要な各クラスタの説明記述については十分な検討が行なわれていない．具体的には，クラスタ内での高頻度語を提示するなど，素朴な戦略を採用しているしている．しかし，この戦略による説明記述提示では有効なナビゲーションがなされないのではないかと考える．そこで，本研究では Scatter/Gather の枠組を援用するが，特に，図 1に示すように，クラスタの説明記述生成ならびに要約生成に焦点を置く．

3.1 文書クラスタリング

検索エンジンによる検索結果の文書を分類するにあたり，文書間距離の定義と，その類似性構造に基づく文書間集合の構造化が必要となる．本研究では文書間距離を

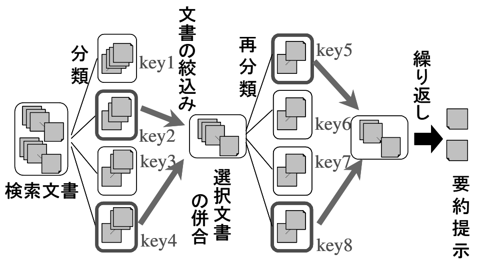

図 1: Scatter/Gather と要約表示による情報ナビゲーション

ユークリッドの距離を用いて次のように定義する．

$$
\begin{aligned}
d(D_i, D_j) &= \sqrt{\sum_k (weight_{ik} - weight_{jk})^2} \\
weight_{ik} &= tf(D_j, w_k) idf(w_k) \\
tf(D_j, w_k) &= \frac{freq(D_j, w_k)}{N} \\
idf(w_k) &= log_2 \frac{N}{df(w_k)}
\end{aligned}
$$

ただし、本稿では語としては名詞に注目し、以下のように定義する。

$freq(D_j, w_k)$	:	文書D_iでの語w_kの出現頻度
$\|D_j\|$	:	文書D_i中の名詞の数
$df(w_k)$	:	検索対象の全文書集合における語w_kを含む文書数
N	:	検索対象の全文書数

クラスタリングアルゴリズムには文書間類似度によって多分岐構造をなし、非階層的な構造を生成する最大距離法 [3] を用いた．具体的なアルゴリズムを以下に示すが，これは Scatter/Gather で採用されている方式とは異なる．

1. 文書集合 D から距離の最も大きい二文書を取り出し，これらを要素とする集合を作成する．これを初期のクラスタ中心の集合 C とする．
2. クラスタ中心集合 C において，クラスタ中心間での最大距離を求める．これを，d_{max} とする．
3. D 中の各文書について，すべてのクラスタ中心との距離を求め，その最小値を既存クラスタからの距離とする．既存クラスタからの距離が最も大きい文書を D から取り出す．この時の距離を d とする．
4. もし，$d \geq \alpha \cdot d_{max}$ ならば, その文書をクラスタ中心集合 C に追加する．もし，$d < \alpha \cdot d_{max}$ ならば，終了．

3.2 クラスタの説明記述の提示

各クラスタに対する説明記述に要求される要件としては，少なくとも以下の二点が考えられる．

1. クラスタ (内の文書) を代表するものであること．
2. あるクラスタについて，他のクラスタとの差異を明らかにするものであること．

クラスタ内の頻度情報のみで説明記述を生成する手法は，上記第一点目に注目したものである．しかし，同手法は他のクラスタでの単語情報は一切考慮されず，他とのクラスタとの差異に関する情報は考慮されないため，クラスタ選択の際に重要となる他のクラスタとの違異点が明確に示されない．そこで本稿では他のクラスタとの差異を測る指標として情報利得比 (informaton gain ratio, IGR) を利用し，この情報利得比を用いたクラスタへの説明記述の生成を検討する．

3.2.1 語の重みとしての情報利得比

この章で述べる情報利得比は本システムにおいて極めて重要な値である．節 3.2.2で述べるクラスタの説明記述の提示のみならず，節 3.3における検索結果文書の要約生成の際にも用いられる．

情報利得比はクラスタの分割構造に対して得られる値で，クラスタの分割の際に毎回算出される．情報利得比とは本来，決定木学習システム C4.5 において属性選択を行うために導入され， C4.5 ではある属性の決定木の分岐におけるテストとしたときに，その属性がどれくらい適切にクラスの出現を予測できるかを表す尺度として用いられている．本研究では属性ではなく，クラスに対応する単語の評価値として情報利得比を用いる．

C_i を C の部分クラスタとするとクラスタ C における単語 w の情報利得比 $gain\_r(w, C)$ は次のように求められる．

$$
\begin{aligned}
gain\_r(w, C) &= \frac{gain(w, C)}{split\_info(C)} \qquad (1) \\
gain(w, C) &= info(w, C) - info_{div}(w, C) \\
info(w, C) &= -p(w|C) \log_2 p(w|C) \\
&- \quad (1 - p(w|C)) \log_2 (1 - p(w|C)) \\
p(w|C) &= \frac{freq(C, w)}{morph(C)} \\
info_{div} &= \sum_i \frac{morph(C_i)}{morph(C)} info(w, C_i) \\
split\_info(C) &= -\sum_i \frac{morph(C_i)}{morph(C)} \log \frac{morph(C_i)}{morph(C)} \\
freq(w, C) &= \text{クラスタ内の語}w\text{の頻度} \\
morph(C) &= \text{クラスタ内の形態素数}
\end{aligned}
$$

$gain(w, C)$ は，クラスタの分割の前後における，語の確率分布に関するエントロピーの減少量を表す．$split\_info(C)$ はクラスタの分割に関するエントロピーである．情報利得比 $gain\_r(w, C)$ は，これらの比として定義される．

上記，情報利得比を語の重みに用いると，出現確率分布について，クラスタの下位分岐構造との整合性が高い語ほど，高い重みが与えられる．すなわち，クラスタを特定できるような語の情報利得比は高い値を示し，クラスタを特定するには至らない語の情報利得比は低い値を示す．

情報利得比はクラスタの分岐毎に重みが計算されるので，情報ナビゲーションの過程で現れる複数の重みについて，その組み合わせ方によって違った重みづけができる．例えば，あるクラスタについてその下位分類を判定するのに特に重要な語を得たいのであれば，そのクラスタについての情報利得比を用いれば良い．一方，それまでの情報ナビゲーションでの文書の絞り込み過程において，どこかで有効であった語に高い重みを与えたいというのであれば，一連のクラスタ分割で得られた値を統合して用いることが考えられる．たとえば，すべてを均等に評価するのであれば，図 2に示すように，式 (2) による和の統合が有効であろう．

$$igr\_sum(w,C) = \sum_{C' \in C_s(C)} gain\_r(w,C') \qquad (2)$$

ただし，$C_s(C)$ は，情報ナビゲーション過程でクラスタ C にいたるまでに現れたクラスタの集合である． なお，情報検索結果文書の場合は，検索された文書と検索されなかった文書の対比も重要であるから，図 3に示すように，この二分割において得られる情報利得比も上記の和に考慮する．以下，あるクラスタに注目したときの情報利得比による語の重みを IGR と呼び，式 (2) による重みを IGRsum と呼ぶ．

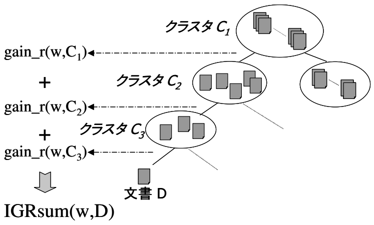

図 2: 情報利得比の統合

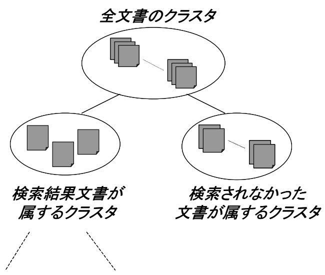

図 3: 最上位クラスタの概念図

3.2.2 情報利得比を利用した説明記述の生成

本稿で想定している，クラスタ構造を提示しながら，部分クラスタを順次利用者に選択してもらうような対話的インタフェースにおいては，情報利得比を利用することにより，各選択点において利用者が注目しているクラスタにおける語の重みを求めることができる．この方法では，ナビゲーションにおいてクラスタを選択する際に重要となる，他のクラスタとの差異を考慮しているため，先ほど述べた 2 つの要件のうち後者を満たすと考えられる．

そこで，この情報利得比による語の重みを，「クラスタを代表している度合を表す尺度」(包含性) を表す語の重みと統合することで，クラスタの説明記述に適する重みを検討する．この包含性に関する尺度の代表例は，Scatter/Gather でも採用されいている TF 値である．また，全文書集合における重要度を表す尺度としては，文書頻度の逆数 IDF 値がある．一方，情報利得比に関する語の重みとしては，前述の IGR,IGRsum がある．そこで， IGR(IGRsum) と TF ,IDF の如何なる組合せによる重みが説明記述 (キーワード) の選定に有効であるかを実験により評価する．

3.3 検索結果文書の要約

現在利用されている WWW 検索エンジン等においても，検索結果に要約が提示される場合がある．しかし，その要約は元文書の最初の数バイトを提示したり，検索要求語の周辺の文を提示するのみで利用者が真に必要とする文書か否かの判断が出来うるほどの十分な品質の要約とは言えない．

我々が想定する情報ナビゲーションシステムにおいても，ある程度文書の絞り込みが行なえたら indicative な要約を提示するわけであるが，以下の点において，検索結果文書の要約が通常の要約とは異なることに注意しなければならない．

- 検索要求文が与えられている．
- 複数の文書が同時に与えられており，ある一度の検索の結果という点において文書間に類似性が見られる．

我々は，後者の情報を元に要約を生成する [4]．

文抽出による自動要約手法を想定した場合，そこで必要となる技術は重要文抽出である．そして，重要文抽出の要となる技術は語の重みに基づく文の重みづけである．よって，先に説明した，説明記述に利用する語の重みづけをそのまま文の重みづけに利用すれば，重要文抽出に基づく基本的な自動要約が実現される．

自動要約においては，単語 w について，式 (2) に基づいて，ある文書に至るまでの情報利得比の合計 $igr\_sum(w,D)$ を求め，これに，文書内頻度 $tf(w,D)$，文書頻度の逆数 $idf(w)$ を組合せることにより、文書 D 中の語 w の重要度 $weight(w,D)$ を得る．ここでは，各値が独立に重要度に寄与するものとし，積を用いて式 (3) のように定義する．

$$weight(w,D) = igr\_sum(w,D)tf(w,D)idf(w) \qquad (3)$$

要約生成において文の重みはキーワードの重要度の平均で表し，式 (4) のように定義する．

$$s_{imp} = \frac{\sum_{w \in keywords(s)} weight(w,D)}{|kewords(s)|} \qquad (4)$$

$$keywords = \text{文}s\text{中のキーワードとなる形態素のリスト}$$

要約文は絶対的な長さにより 150 形態素と定め，下記の評価実験を行なった．

4 各種語の重みを用いた説明記述の比較実験

4.1 実験

被験者 3 人 (大学院生 2 人，大学生 1 人) に本システムを用いて特定の情報を探索するタスクを行なってもらった．質問 (例えば「日本三大祭とは何か?」など[1]) が与えられるので，被験者は自分で本システムに検索キーワードを与え，情報検索システムが返した上位 200 件の記事を対象にして探索を開始する．検索対象文書は毎日新聞 98 年ならびに 99 年の全記事である．

検索結果文書は最大距離アルゴリズム ($\alpha = 0.8$) によりクラスタリングされ幾つか (10 程度) の文書グループに分類される．クラスタリングの度に被験者に提示される情報は，各文書グループ毎に，所属文書数，ならびに，説明記述 (語の重みで順位づけした上位 10 位までの語 (複合語も含む) をその順番でならべたもの) である．提示された情報だけを基にして，被験者は関連があると思われる文書グループを選択 (複数可) し，次の分類を行なう．上記過程を繰返した後，必要な文書が十分絞りこめた場合 (20 文書以下に設定) は，要約提示を要求することができる．その要約を読むことにより，最初の質問の答を回答する．語の重みとしては，IGR，IGRsum，TF(クラスタ内頻度)，IDF(全文書集合を対象にした値)，TF·IDF，IGRsum·TF，IGRsum·TF·IDF，IGRsum·IDF を各々設定してタスクを遂行してもらった．

答を得るまでの操作ステップ数 (タスク遂行時間に対応)，ならびに，実験の後の被験者への提示される各種説明記述の内容的な傾向，特徴を聞き，各種語の重みによる使い勝手の良さ悪さに対する定性的な情報を得た．

4.2 評価

まず，IGR と IGRsum を比較した場合，明らかに IGRsum の方が有効であるという意見であった．IGRsum では，過去の選択過程をすべて考慮していることに加え，検索文書と検索されなかった文書の間の対比に関する情報が含まれているために，検索要求に関する関連する語が提示されるので，クラスタ選択を行なう上で有利となると考えられる．

また，TF もしくは TF·IDF による説明記述よりも，IGRsum·TF，もしくは，IGRsum·TF·IDF による説明記述の方が答を含む文書への到達が早くなるという傾向があった．すなわち，説明記述としては，IGRsum を組み合わせた重みの方が文書の絞り込みが適切に行なえることを示すものである．

TF を含めない手法，すなわち，IGR，IGRsum，IGRsum·IDF に関しては，特にナビゲーションの初期の段階において，質問に関連する語が把握できず，クラスタの選択に困る場合が多かったという意見が主流であった．すなわち，TF はクラスタ内の文書の特徴を表すものであるので必要な尺度と考えられる．

IDF についていえば，一つの文書に集中している度合であるので，人名や専門用語のような語の重みが高くなりやすい．これを重みに加えた場合，被験者が事前に持つ質問に対する知識の量によって，有益か否かが分かれる傾向が見られた．

以上より，小規模な実験による定性的な評価ではあるが，TF· IGRsum，もしくは TF· IDF· IGRsum による説明記述がナビゲーションに適していると考えられる．この二つの手法違いは IDF を考慮しているかどうかであるが，これは先ほど述べた理由により利用者の背景知識によるところが大きい様である．

[1] NTCIR3 QAC Dryrun の質問を利用させて頂いた．

5 まとめと今後の課題

本稿では情報利得比を語の重みとする説明記述ならびに文章要約に基づくナビゲーションシステムについて提案し，小規模実験による定性的な評価を行なった．今後，更に詳細な定量的な評価を行なう予定である．

参考文献

[1] A. Tombros and M. Sanderson. 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*, pp.2–10 (1998)

[2] Douglass R. Cutting , David R. Karger , Jan O. Pedersen , John W. Tukey. Scatter/Gather : A Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of the Fifteenth Annual International ACM SIGER Conference*, pp.318–329 (1992)

[3] 長尾真. パターン情報処理. 電子通信学会大学シリーズ, コロナ社, pp.116–117 (1983)

[4] 菊池美和, 吉田和史, 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. 言語処理学会第 7 回年次大会発表論文集, p.189-192 (2001)