情報検索ナビゲーションにおける ユーザへの提案機構と可視化インタフェース

伊藤直之 † Nikolay Elenkov‡ 森辰則 ††

† 横浜国立大学 大学院 環境情報学府 ‡ 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: {spud,nick,mori}@forest.eis.ynu.ac.jp

1 はじめに

クラスタリングに基づくナビゲーション過程から得られるクラスタへの適合評価を利用して，検索意図を自動的に学習する手法が提案されている．江口ら [?] は，検索過程において，クラスタ選択によるフィードバック情報を利用して検索質問を漸次的に拡張し，適応的クラスタリングに利用している．しかし，拡張された検索質問による再検索を実行する場合，ナビゲーション過程で絞り込まれた文書群は保持されず，再度，文書の絞り込みを行わなくてはならないといった問題点を挙げることができる．

本稿では,Scatter/Gather[?] に基づくナビゲーション過程において，クラスタ選択によりユーザが絞り込んだ文書群は保持しつつ，拡張された検索質問を再検索に利用し，システムからの推薦文書群を提示する手法を提案する． また，同システムで使用する可視化インタフェースについても述べる.Hyperbolic tree 等により，クラスタの構造や距離を可視化することで，より直観的なブラウズを支援する．

2 提案手法の概要

Scatter/Gather[?] は，動的なクラスタリングにより対話的に文書を絞り込むものである．まず，検索結果文書が決められた数のクラスタに分類され，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群が付与される．利用者はその説明記述をもとに，検索意図に適合していると思われるクラスタ群を選択する．選択されたクラスタ群は併合され再分類される．以上の処理を繰り返すことにより文書集合を絞り込んでいく (図??).

また，ユーザがシステムに検索文書への適合性判定を伝えることにより，検索効率を向上させる適合性フィードバックの手法が提案されている．得られた適合性判定を用いて，検索質問に新たな索引語を追加し，索引語の再重み付けを行うことで，検索質問を拡張する．拡張された検索質問により再検索を行うことで，ユーザの検索意図に，より適合した文書群を得ることができる．しかし，適合性判定に伴うユーザの負荷が生じる．

本稿では,Scatter/Gather の過程におけるクラスタ選択を，適合性フィードバックにおけるユーザの適合性判定ととらえることでユーザの興味を学習し，検索意図に適合した推薦文書を提示する手法について述べる．クラスタ選択に基づいて適合性フィードバックが適用されるので，ユーザに負荷をかけずに，自動的に検索意図を学習できる．ユーザのクラスタ選択を尊重するため，推薦文書は，絞り込んだ既存文書のクラスタ群とは独立して分類し提示する．ユーザは既存文書のクラスタに加え，新たに提示された推薦文書のクラスタを選択することができる (図??)．また，提案する推薦文書をよりユーザの検索意図に近付けることを目的に，推薦文書の選択においても，適合性フィードバックを行う．

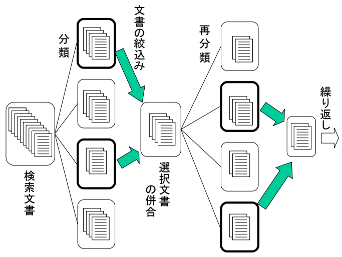

図 1: Scatter/Gather の概念図

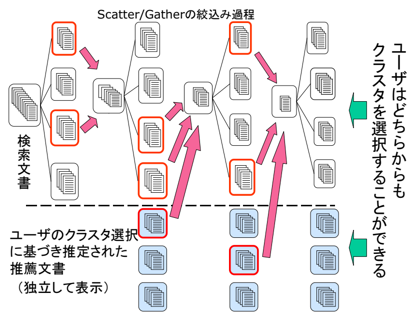

図 2: Scatter/Gather 過程での推薦文書の提示

3 クラスタ選択に基づく適合性フィードバックと関連文書の提案

3.1 説明記述キーワードと文書との類似度の利用

従来のクラスタ選択に基づく適合性フィードバックでは，適合判定されたクラスタ内の全文書を適合文書として一律に扱う．不適合と判定されたクラスタについても同様である．しかしナビゲーション過程でのクラスタ選択の際にユーザが閲覧するのは，あくまでもクラスタに付与された説明記述キーワード群であり，文書一つひとつではない．そのため，クラスタ内の全文書を一律に扱う方法では，ユーザの選択意図を的確に捉えられないと考えられる．そこで，クラスタに対して付与された説明記述

キーワードと, そのクラスタに含まれる文書との類似度を用いて, 文書に重み付けを行った上でフィードバックする. ユーザがクラスタを選択する際に閲覧した説明記述キーワードを含む文書の重みを大きくすることで, より的確に検索意図を捉えるためである. あるクラスタ C に付与された説明記述キーワード群を, キーワードに対応する成分を 1 とした文書ベクトル k_C で表す. また, クラスタ C に含まれる文書の文書ベクトル d を, 文書内の語の重みを要素とするベクトルとする. この時, 各文書とクラスタ C の説明記述キーワードとの類似度を各ベクトルの内積により求めることができる (式 (??)).

$$sim(d, k_C) = d \cdot k_C \tag{1}$$

式 (??) の類似度を用いて, クラスタ C に含まれる文書ベクトル d に対して重み付けしたものを d^k とする (式 (??)).

$$d^k = sim(d, k_C)d \tag{2}$$

この重み付けされた文書ベクトルを用いてフィードバックを行う. なお, クラスタ説明記述の生成には, 語の重みとしての情報利得比 (IGR) と $tf \cdot idf$ 値を組み合わせた値によるキーワード選択法 [?] を用いる.

3.2 クラスタへの適合性判定に基づく質問拡張

ベクトル空間モデルを用いた Rocchio 式 [?] を修正した式 (??) により, 検索質問の拡張を行う. 拡張された検索質問を用いて再検索を行い, 得られた文書群をクラスタリングし, 推薦文書としてユーザに提示する.

$$\begin{aligned} q_{n+1} = \hat{q_n} \quad &+ \quad \frac{\alpha}{|\cup_{C_r \in RC} C_r|} \sum_{C_r \in RC} \sum_{d_i \in C_r} \hat{d_i^k} \\ &- \quad \frac{\beta}{|\cup_{C_n \in NC} C_n|} \sum_{C_n \in NC} \sum_{d_j \in C_n} \hat{d_j^k} \\ &+ \quad \frac{\acute{\alpha}}{|\cup_{Cs_r \in RCs} Cs_r|} \sum_{Cs_r \in RCs} \sum_{d_l \in Cs_r} \hat{d_l^k} \\ &- \quad \frac{\acute{\beta}}{|\cup_{Cs_n \in NCs} Cs_n|} \sum_{Cs_n \in NCs} \sum_{d_m \in Cs_n} \hat{d_m^k} \end{aligned} \tag{3}$$

ナビゲーション過程での絞り込みステップ n においての検索質問ベクトルを q_n とすると, ユーザのクラスタ選択に基づく適合性フィードバックにより, 次ステップでは検索質問ベクトルが q_{n+1} へと漸次的に拡張され, 推薦文書の検索の利用される. ここで $d_i^k, d_j^k, d_l^k, d_m^k$ は, それぞれ, 既存文書クラスタ中適合と判定されたクラスタに属する文書, 不適合と判定されたクラスタに属する文書, 推薦文書クラスタ中適合と判定されたクラスタに属する文書, 不適合と判定されたクラスタに属する文書の各文書ベクトルを, 式 (??) により重み付けしたものである. また,$\hat{d_i^k}, \hat{d_j^k}, \hat{d_l^k}, \hat{d_m^k}$ は,$d_i^k, d_j^k, d_l^k, d_m^k$ を, それぞれベクトルの長さが 1 になるように正規化 (コサイン正規化) したものである. RC は, 既存文書において, ユーザが適合文書を含むことを期待して, 適合と判断したクラスタ C_r の集合であり,NC は, 不適合と判断したクラスタ C_n の集合である.RC_s は, 提案された推薦文書から, ユーザが適合文書を含むことを期待して, 適合と判断したクラスタ Cs_r の集合であり,NC_s は, 推薦文書から不適合と判断したクラスタ Cs_n の集合である.α,β,$\acute{\alpha}$,$\acute{\beta}$ は 0 以上の定数であり, それぞれ, 既存文書クラスタ中適合と判定されたクラスタに属する文書, 不適合と判定されたクラスタに属する文書, 推薦文書クラスタ中適合と判定されたクラスタに属する文書, 不適合と判定されたクラスタに属する文書をどの程度重要視するかを表している. 本稿では, ユーザが適合と判断しなかった文書クラスタに関しては全て不適合と判断されたとして, フィードバックを行う. フィードバックの際に使用する語の重みには $tf \cdot idf$ 値を用いる. なお, ユーザによる推薦文書クラスタ群からの選択が行われなかった場合は,$\acute{\alpha}$,$\acute{\beta}$ ともに 0 とする.

4 クラスタリングに基づく検索結果の閲覧のための可視化インターフェース

4.1 本インターフェースの目的

Scatter/Gather の過程において, ユーザは自らの検索意図に適合するであろう文書クラスタの選択を行うが, クラスタについての説明記述キーワードを提示するだけでは, ユーザのクラスタ選択の支援には不十分であると我々は考える. そこで, クラスタリング結果を可視化して表示するインタフェースを構築した. クラスタの構造と属性を直観的な形式で可視化し, より使いやすいインターフェースの構築を目標とした.

4.2 可視化手法の概要

クラスタの木構造の可視化について, 直観的な構造の把握と, 効率的な画面領域の使用という観点から, 以下の 3 種の可視化手法を採用した.

- Hyperbolic tree
- Tree-maps
- GEM アルゴリズム

Hyperbolic tree は, 木構造の, 最も直観的な可視化を実現する手法のひとつである [?]. 双曲面にレイアウトされた木構造を円盤に射影するので, 注目している箇所を円盤の中心に配置することでフォーカスされ, それ以外の部分は円盤の端で小さく表示される. この機能により, クラスタ構造の注目する部分だけを大きく表示することができ, 効率的な画面領域の使用が可能となる.

Tree-maps[?] では, 各クラスタは長方形として表示され, その面積は該当するクラスタの大きさを表している. Tree-maps の適用でクラスタの大きさと構造の直観的な認識を容易にすることができる.

クラスタ間距離 (または類似度) を 2 次元平面で可視化する代表的な手法として, 自己組織マップとバネモデルに基づく手法がある. 自己組織マップはニューラルネッ

トワークに基づくため, 実行速度が遅い. 一方, バネモデルは比較的簡単に実現できるが, システムが収束するまでに時間がかかる場合が多いので, インタラクティブなアプリケーションには向いてない. そこで, バネモデルをローカル温度, 重心力, 振動等で拡張した, 収束速度の速い GEM アルゴリズム [?] を採用した.

ユーザはナビゲーション過程のなかで, クラスタリング結果の表示方法を自由に切り替えることができる. また, 絞り込みを進めるごとに, 推薦文書のクラスタ群が, 既存文書のクラスタ群と同時に, 独立して画面表示される. 図??は,Hyperbolic tree によりクラスタ構造を表示した例である. 既存文書クラスタ群と推薦文書クラスタ群が, 上下に区別され, 独立して表示されている. ユーザは, 双方のクラスタ群から, クラスタ選択を行い, 絞り込みを進めることが可能である.

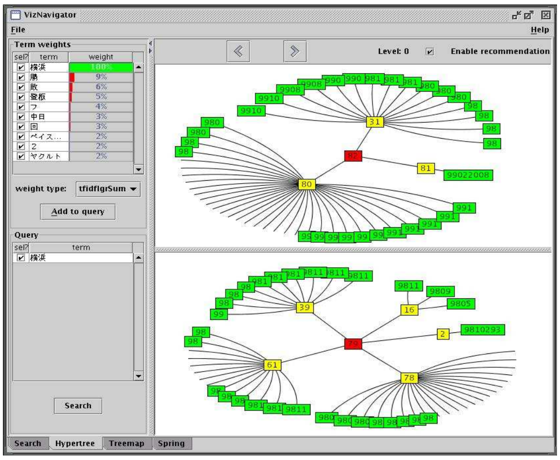

図 3: 既存文書クラスタ群 (上段) と推薦文書クラスタ群 (下段) の表示例 (Hyperbolic tree を用いた場合)

5 評価実験

本稿で述べた提案機構を組み込んだ情報検索ナビゲーションシステムについて評価実験を行った.

5.1 評価方法

NTCIR テストコレクション[1]を用いて, 提案手法を評価検討する. 検索質問としては, 各検索課題に付与されている課題タイトルのみを使用した. 各質問をユーザが与えたものとみなし, ユーザによるクラスタの選択を各種, 想定した上で, その際に提示される推薦文書の正解文書に対する適合率・再現率や, 推薦文書に含有される新規の正解文書 (新文書) の数を調べた.

初期検索を行ったときの上位 N 文書[2]が, クラスタリングされ, その中からユーザはクラスタを選択し, 絞り込みを行う. 利用者にクラスタ構造を提示した際, 分割された構造が利用者にとって見やすいであろうと判断し, クラスタリングによる分割数を 9 とした. クラスタが選択され, 再クラスタリングが実行される度に, 推薦文書として N_R 文書[3]が提示される. 適合性フィードバックのパラメータには, 従来多く用いられてきた [?],$\alpha = 2.0, \beta = 0.5$ を用いることにし, 推薦文書クラスタの選択によるフィードバックのパラメータについても $\acute{\alpha} = 2.0, \acute{\beta} = 0.5$ に設定した.

クラスタリング結果において, 含有する正解文書数が最も大きいクラスタをベストクラスタ (Best Cluster) と呼ぶことにする. また Scatter/Gather の一段階を step という単位で表す. すなわち初期クラスタリングの結果からクラスタを選択し, 併合, 再クラスタリングが行われた段階が step1 であり, 以降, 再クラスタリングを実行するごとに step2,step3 と順次表すことにする.

5.2 既存文書からのクラスタ選択により提示される推薦文書の精度評価

理想的な状況として, ユーザがクラスタリング結果から常にベストクラスタのみを選択して絞り込みを行った場合と, 正解文書の含まれるクラスタを全て選択して絞り込みを行った場合を想定し, 各 step で提案機構により提示される推薦文書の精度を比較した. なお, 既存文書クラスタのベストクラスタのみを選択し続けた際に, 選択された文書数がクラスタ分割数 9 を下回った場合は, その検索課題については検索終了とみなし, 次のステップに進まない. 表??は,step6 までのナビゲーション過程において, それぞれの step で提示される推薦文書の平均適合率を示したものである.step2 以降においては, 初期検索と比較して, 推薦文書が高い適合率をもつことがわかる. また, クラスタ選択による文書群の絞り込みを進めるごとに, 適合率は大きくなる傾向にある.Scatter/Gather の過程で適切なクラスタ選択が繰り返し行われると, そのクラスタ内の文書による適合性フィードバックにより検索効率が漸進的に向上する点は注目に値すると言え, ユーザの検索意図を効果的に学習できていることがわかる. また, ベストクラスタのみを選択したときと, 適合クラスタを全て選択したときを比較すると, 前者の方が, より適合性の高い文書群を提案できていることがわかる. ベストクラスタのみを選択した場合には, 適合していると判定された文書全体に対する正解文書の比率が大きいため, クラスタ選択過程において効果的にユーザの興味を学習できたものと考えられる. 特に, ユーザが適切なクラスタを 1 つだけ選択するだけで良いという点が興味深い. 表??は, 初期検索質問で検索されなかった文書を, どれだけ推薦文書として提案できたかを示したものであるが, クラスタ選択からユーザの興味を学習し, 初期検索で

[1] NII-NACSIS Test Collection of IR Systems
NTCIR-3 の言語横断情報検索 (CLIR) 用テストコレクションの中から, 日本語の検索対象のみを用いた.

[2] 実験では $N = 200$ とした.

[3] 実験では $N_R = 200$ とした.

表 1: 初期検索結果と推薦文書の平均適合率

	初期検索	step1	step2	step3	step4	step5	step6
A	0.251	0.234	0.314	0.344	0.321	0.399	0.418
B	0.251	0.212	0.264	0.279	0.315	0.362	0.328

A : 既存文書クラスタ中, ベストクラスタのみを選択
B : 既存文書クラスタ中, 正解文書を含むクラスタを全て選択

表 2: 推薦文書に含まれる新文書

	step1	step2	step3	step4	step5	step6
Rel	1249	1341	1153	930	800	422
Rel (old)	872	899	808	668	601	319
Rel (new)	377	442	345	262	199	103
Novelty	0.302	0.330	0.299	0.282	0.249	0.244

Rel : 推薦文書中の正解文書数
Rel (old) : 推薦文書中, 初期検索で検索されたもの
Rel (new) : 推薦文書中, 初期検索で検索されなかったもの
Novelty : Relevant(new) / Relevant

は探し出せなかった文書を提案できていることがわかる.

5.3 推薦文書からのクラスタ選択が行われた場合

推薦文書クラスタからも, クラスタ選択を行った場合についても調べた. その結果, 各 step において, 既存文書クラスタ, 推薦文書クラスタの両方から, それぞれのベストクラスタを選択していった場合に提示される推薦文書は, 既存文書クラスタのベストクラスタのみを選択した場合に提示される推薦文書と比較して, 低い適合率を示した. ベストクラスタを常に選択していく場合においても, 初期クラスタリング結果に対する一回目のクラスタ選択によるフィードバックによって提示される推薦文書は, 初期検索と比べても適合率の低いものとなっていることから, 適合すると判定された文書全体に対する正解文書の比率が低いと, 効果的な学習が行われないと考えられ, 全く絞り込みが行われていない推薦文書からのクラスタ選択によるフィードバックからは, ユーザの検索意図がうまくとらえられなかったと考えられる.

しかし, 本稿で述べた推薦文書群に望まれる内容として考えられるのは, 適合文書を数多く含むとともに, ユーザが与えた検索質問では検索されなかった文書をどれだけ多く含むかということである. そこで, 推薦文書中, 初期検索で探し出されなかった文書の数を調べてみると, step2 以外では, 既存文書クラスタと推薦文書クラスタの両方のベストクラスタを選択した場合の方が若干大きい値となっていた. ユーザが与えた検索質問では検索されなかった文書をどれだけ多く提示できるかという, 提案機構に望まれる能力の点から考えると, ユーザによる推薦文書クラスタの選択からもフィードバックを得ることは有意義であるといえる.

同様に, 初期検索結果の絞り込みを数 step に渡り行った後, 提示された推薦文書クラスタを選択したときの, 次 step における推薦文書群についても調べたが, 大きな検索精度の向上には結び付かなかった.

ただし, 本節での評価は初期検索質問に対する関連度により行っているが, 推薦文書に対する評価は, 本来, 検索要求が変遷する状況で行うことが望まれる. ユーザが検索質問で記述しきれなかった情報要求をナビゲーション過程で明らかにすることが本稿で述べた提案機構の重要な役割であるからである.

6 まとめと今後の課題

本稿では, クラスタリングに基づく情報ナビゲーションシステムにおいてブラウジング過程から得られるクラスタへの適合評価を利用してユーザの検索意図に適合した推薦文書群を提案する機構について述べた. また, ナビゲーションのためのグラフィカルユーザインタフェースを構築した.Scatter/Gather によるナビゲーション過程で適切なクラスタ選択が繰り返し行われると, そのクラスタ内の文書による適合性フィードバックにより, 推薦文書の検索効率が漸進的に向上することを示した. また, 推薦文書の選択からもフィードバック情報を得て, 検索効率の向上を試み, 推薦文書中の新文書含有率において若干の向上が見られた.

今後は, ユーザの検索要求がナビゲーション過程で変遷する状況を想定しての, 本システムの有効性についての評価を行う予定である.