データベースシステム　118－2
情　報　学　基　礎　54－2
（1999．5．17）

文書クラスタリングを利用した検索質問展開手法の開発と評価

新田 清，蓬莱 尚幸，園部 正幸
{kiyoshi, horai, sono}@flab.fujitsu.co.jp
（株）富士通研究所

要約: 我々は文書クラスタリングエンジンを用いて，検索質問展開に文書クラスタリング技術を適用した．文書クラスタリングを利用した検索質問展開手法(以下文書クラスタ QE)では，予備検索の結果から数個の最良の文書クラスタを選んで検索質問展開を行う．文書クラスタ QE は，予備検索の結果上位数件(例えば 1,000 件)を用いるような単純な検索質問展開よりも良い結果を生成した．さらに文書クラスタ QE は質問群の差に対して高成績で安定した手法であり，文書群の性質や連携する検索エンジンの特性に対してもフレキシブルであった．

Clustering-based Query Expansion: Development and Evaluation

Kiyoshi Nitta, Hisayuki Horai, Masayuki Sonobe
FUJITSU LABORATORIES LTD.

Abstract: We developed a clustering-based query expansion (QE) and implemented it using a text-clustering engine. Clustering-based query expansion uses the top N best document clusters from the top M documents instead of just using the top N documents. Clustering-based QE produces better results than simple query expansion based on passage retrieval. The clustering-based QE can be applied to wide variety of query sets with good performance, and it is flexible in being applied to various sets of documents and cooperating with various search engines.

1. はじめに

情報検索システムには統計的手法に基づくものの他にも認知的手法に基づくものなどの研究が進められている[1, 2]が，現在，実際に用いられている多くの全文検索システムは統計処理を原理とする検索エンジンである．この種の検索システムの性能を向上させるために様々な手法が研究されてきた．検索質問展開手法もその一つである．

全文検索システムを用いて検索を行う場合，ユーザは探したいものを表現する語をシステムに入力する．この検索結果にユーザが満足しない場合，結果として出力された文書の内容を参考にしながら，入力する語を修正し再検索する．このように検索システムへの入力を修正するプロセスを検索質問展開という．ユーザの利便性向上のために，このプロセスの処理中に人手が全く介在しない自動化についても研究されている．本稿ではこのような自動的検索質問展開手法（automatic query expansion method; 以下，単に QE と呼ぶ.）について議論する．

予備検索における上位文書から単語を取り出して検索語にするような展開を自動的に行う手法は従来ひろく行われている[3]（以下，従来型 QE と呼ぶ.）が，この手法には次の問題が指摘されている；

(P-1) 予備検索の結果で正解文書が下位に集まるような場合に性能が落ちる

従来型 QE では予備検索における固定された上位 n 件を用いて処理され，精度を上げるために n を小さく抑える．そのため展開に用いない文書が

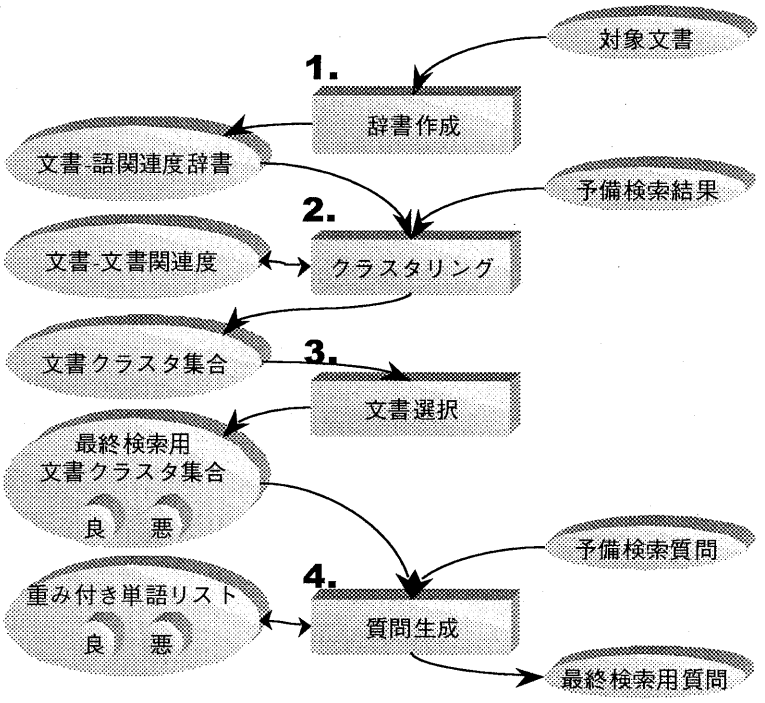

図 1 文書クラスタ QE の処理概要

多くなり，その中に含まれる語が QE に使われなくなる．

また一般に QE を開発するとき，a) 連携する検索エンジン，b) 対象文書の種類，c) 検索質問の傾向，に依存して作られる．このことからは次の問題が生じる；

(P-2) a), b) が変更された場合の対応に開発工数が多くかかる

(P-3) c) が変化した場合に性能が維持できない

以上の問題を改善することを意図して，予備検索で取得した文書をクラスタリングして選択し，さらにそれらの文書から統計手法により関連語を選び出す手法（**文書クラスタ QE**）を提案する．

以下，第 2 章では文書クラスタ QE の概要を説明し，第 3 章では文書クラスタ QE を実装した具体例を関連技術と共に示し，性能評価についても述べる．考察を第 4 章で行う．

2. 文書クラスタ QE

第 1 章で述べた 3 つの問題を解決するために我々は文書クラスタ QE を開発した．本章ではその概要を述べる．**図 1** に文書クラスタ QE の概要を示した．ここで矩形は処理の段階を示し，楕円はデータを示す．矢線はデータの流れを示す．右手列の楕円は文書クラスタ QE が外部（ユーザや検索エンジン）と連携するときに媒介するデータであり，左手列の楕円は文書クラスタ QE が内部使用するデータである．

文書クラスタ QE は，1）辞書作成処理，2）文書のクラスタリング処理，3）文書の選択処理，4）最終検索用質問の生成処理の 4 段階から構成される．辞書作成処理は時間がかかり，検索質問とは独立して処理できることから，検索対象文書が定まった段階で前処理として行う．残りの処理は質問毎に行い，複数の質問を一度に処理するときは繰り返し実行する．

辞書作成処理では検索対象文書から文書-語関連度辞書を作成する．この処理では語の重みメジ

ャーと正規化方法を変更することで傾向の異なる関連度辞書を作成できる．これらの変更に関する詳細については後述する．

クラスタリング処理では予備検索で取得した文書のクラスタの集合を生成する．以下，それぞれのクラスタを**文書クラスタ**，文書クラスタの集まりを**文書クラスタ集合**と呼ぶ．この処理では文書-文書関連度を計算するための関連メジャーや，クラスタリング手法を変更することで傾向の異なる文書クラスタを生成することができる．これらの変更に関する詳細については後述する．

文書選択処理では文書クラスタ集合を用いて質問展開用文書を選択する．質問展開用文書として正解文書が多いという意味の「良い」文書クラスタ集合と，少ないという意味の「悪い」文書クラスタ集合を生成する．

文書選択処理は文書クラスタ集合，文書クラスタ，文書（これら3つをまとめて対象と呼ぶ）の選択を複数組み合わせた処理である．対象の選択では，まず，予備検索における文書順位（以後，RPSと呼ぶ；Rank of Pilot Search）などの対象の属性に基づいて対象を評価する．対象の属性を表1に示す．次に，評価の結果（評価値）を用いて対象を選択する．評価値を用いた選択方法には，評価値の大小に基づいて対象に順序を付けて上位数件を選択する方法と評価値がある範囲である対象を選択する方法がある．

質問生成処理では，良い文書クラスタ集合と悪い文書クラスタ集合から最終検索用質問を生成する．まず，良い文書クラスタ集合と悪い文書クラスタ集合より，それぞれ文書-語関連度が高い複数の単語を選択し，文書-語関連度を重みとした重み付き単語リストを生成する．次に，重み付き単語リストと予備検索質問（以後，QPSと呼ぶ；Query for Pilot Search）を配合し，最終検索用質問とする．この配合において，悪い文書クラスタ集合から生成された重み付き単語リストは良い文書クラスタ集合から生成された重み付き単語リストに含まれる不適切な単語の重みを抑制するために用いる．この抑制は最終検索の精度を高める上で重要である．質問生成処理では文書-語関連度を計算するための関連メジャーを変更することで異なる傾向の重み付き単語リストを生成できる．また良い・悪い重み付き単語リストと予備検索質問の配合方法を変更することで，異なった最終検索用質問を生成することができる．予備検索質問と最終検索用質問の例を図2に示した．

表1 文書選択の評価対象と評価属性

評価対象	評価属性
文書	順位，スコア，文書が属すクラスタの属性
文書クラスタ	平均値，再頻値，中間値，最高値，最低値，文書数，文書クラスタが属す文書クラスタ集合の属性
文書クラスタ集合	平均値，再頻値，中間値，最高値，最低値，文書クラスタ数

3. 具体例：TREC

前章で概要を示したクラスタリング QE を，TREC7 ad hoc task 用に実装した（実装したシステムをTREC用システムと呼ぶ）[4,5]．TREC7 ad hoc task での検索対象の文書は英文の新聞や経済分野の報告書約 55 万件，2 ギガバイトある．主催者は質問（例は図 2 上部を参照．）を 50 問用意し，それぞれの質問に対して正解を用意した．

本章では TREC 用システムの実装に関連する技術，そのために行った処理フローの選択，パラメータのチューニング，そして TREC による検索システムとしての総合的な評価について述べる．

3.1 要素技術と検索エンジン

文書クラスタ QE はクラスタリング技術を要素として利用した．クラスタリング技術には，文書や語の間の関連度を統計的に算出するための Keyword Associator[6,7]（以下，KA と呼ぶ．）と，KA と連携してクラスタリング機能を提供する KAcluster を利用した．QE は検索システムを構成するために検索エンジンと連携することが必要である．TREC 用文書クラスタ QE では，大量文書の高速全文検索機能を提供する検索エンジン Terass[8] と連携した．これらは全て富士通研究所で開発された．以下それぞれについて特徴をさらに述べる．

3.1.1 Keyword Associator

Keyword Associator は，文書と語の間の関連度を計算し辞書に格納する処理や，その辞書を用

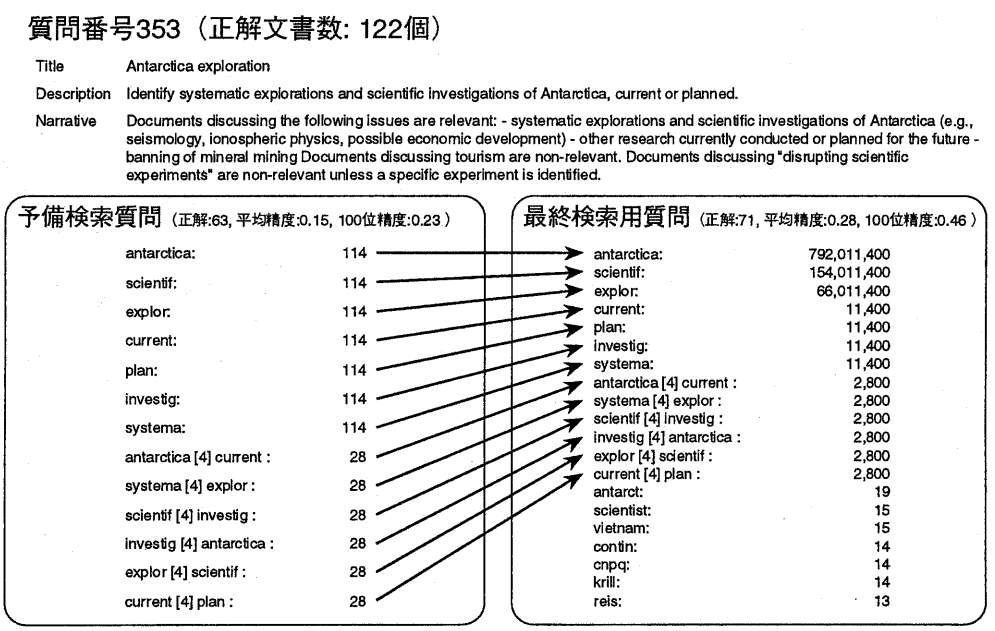

図 2 TREC における質問と文書クラスタ QE 入出力の例

いた検索等の操作を行う．KA は豊富な機能を提供し，コマンドラインまたは KA-batch スクリプト言語により機能を制御できるという特徴を持っている．この KA-batch を備えた KA を要素技術として利用したことにより，文書クラスタ QE はフレキシブルなシステムとなった．

文書クラスタ QE では，KA が提供できる多数の機能のうち，

(KA-1) 文書群を入力し，それらに関連する複数の語を文書群との関連度と共に出力する

(KA-2) 文書群を入力し，それらの間の関連度をマトリックスとして出力する

という二種類の機能を利用した．(KA-1)では，語の重みメジャーと正規化方法を指定する．語の重みメジャーは，LtLnu[9]，tf，idf，等から選択でき，正規化方法は，適用せず，語ベクトル和=1，語ベクトル 2 乗和=1，文書ベクトル和=1，などから選択できる．(KA-2)ではさらに関連メジャーを指定する．関連メジャーは，内積，Dice coefficient，Cosine coefficient，Jaccard coefficient[10]から選択できる．

3.1.2 KAcluster

KAcluster は KA から関連度を受け取り，要素（文書または語）をクラスタリングする．クラスタリング手法は，階層的凝集型手法(HACM)[11,12]である Single Link 法，群平均法，Median 法，Centroid 法，Ward 法から選択できる．KAcluster は階層的凝集型手法が生成する全ての階層のクラスタを出力する．

3.1.3 Terass

全文検索エンジン Terass は大量の文書情報を高速に検索する．検索速度を向上させるために独自のインデックス圧縮方法を採用しており，ギガバイトを超える大容量文書の検索を十分実用的に行う性能を持つ．この高速性によりチューニングの際には多数のパラメータを短期間で試すことができ，文書クラスタ QE の性能を向上させることに大変役立った．

3.2 処理フローの選択

文書クラスタ QE が性能を発揮するためにはチューニング作業が必要である．チューニング作

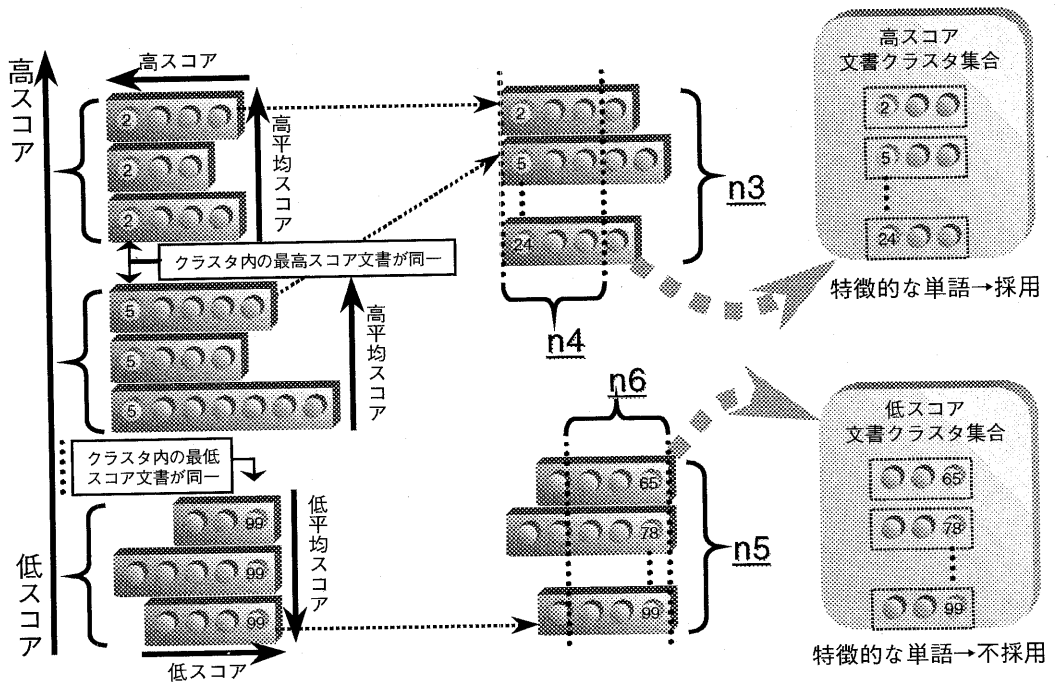

図 3 文書選択方法

業は 1) 処理フローの選択, 2) パラメータのチューニングからなる. 本節では実際に選択した処理フローとその選択理由を述べる.

3.2.1 辞書作成

KA を用いて辞書作成を行った. 具体的には, KA の(KA-1)の機能を利用した. 語の重みメジャーとしては LtLnu メジャーを, 正規化方法としては語ベクトル 2 乗和=1 を選択した. 他のほとんどの処理がこの辞書の影響を受けることは明らかであるため, このパラメータは最初に決定した. 他の処理のパラメータを暫定的に固定して重みメジャーの種類を複数試したところ, 比較的良い TREC スコアを出したためこの選択となった.

3.2.2 文書のクラスタリング

KA および KAcluster を用いて予備検索結果(上位 1,000 件) をクラスタリングした. 具体的には, KA の(KA-2)の機能を利用した. 関連メジャーとしては内積を用いた. 得られた予備検索結果間の関連度マトリックス値を, KAcluster により Word 法でクラスタリングした.

文書クラスタ QE では正解文書を多く含むクラスタが多数できあがるようなクラスタリング手法が望ましい. Ward 法以外の凝集型クラスタリング手法では鎖状クラスタができやすく(chaining effect), 非正解文書を多く含む巨大クラスタが形成されやすい. このため比較的バランスのとれたクラスタを生成する Ward 法を選択した.

3.2.3 文書の選択

得られた予備検索結果のクラスタリングから,「良い文書クラスタ集合(GCS)」と「悪い文書クラスタ集合(BCS)」の 2 個の文書クラスタの集合(文書クラスタ集合)を生成する.

次に, GCS と BCS の生成方法について述べる. 図 3 に本処理の概要を示す.

1. 適切な大きさ(要素数)を持つ文書クラスタの選択 : 要素が n1 個未満の文書クラスタおよび n2 個より多い文書クラスタを取り除く.
2. GCS の生成
   - 2.1 文書クラスタの評価 : 各文書クラスタについて, 1)RPS が最も良い文書をその文書クラスタの代表文書(REP)とし, 2)要素である文書の RPS の平均値(AVE)を計算する.
   - 2.2 文書クラスタ集合の生成 : REP が同じ文書クラスタを一つの文書クラスタ集合にまとめる

ことで，複数の文書クラスタ集合を生成する．

2.3 文書クラスタ集合を用いた文書クラスタの選択：各文書クラスタ集合からAVEが最も良い文書クラスタを選択することで，1個の新たな文書クラスタ集合を生成する．生成された文書クラスタ集合の要素のREPは全て異なる．

2.4 良い文書クラスタの選択：生成された文書クラスタ集合に対して，その要素である文書クラスタのうちREPのRPSがn3より良いもののみを残し，それ以外の文書クラスタを文書クラスタ集合から取り除く．

2.5 良い文書の選択：文書クラスタ集合の要素である各文書クラスタについて，RPSがn4より良い文書のみを残し，それ以外の文書を取り除く．得られた文書クラスタ集合がGCSである．

3. BCSの生成：ステップ2と同様にBCSを生成する．ただし，ステップ2における「良い」，GCS，n3，n4は，それぞれ，「悪い」，BCS，n5，n6に置き換える．

後段の検索質問生成処理を成功させるためには，文書クラスタ集合から「適度な」多様性を持つ文書クラスタを選択することが重要である．もしAVEのみを用いて文書クラスタを選択した場合，選択される文書クラスタの多様性は制限され過ぎる．ここでは，文書クラスタ集合を用いた文書クラスタの選択（ステップ2.3）を導入することでAVEのみを用いる選択よりも高い多様性を実現した．

3.2.4 検索質問の生成

次に，GCSとBCSとQPSを用いて，最終検索用質問を生成する方法について述べる．ここでは，検索質問は重みが付いた単語の集合とする．これ以降，重みが付いた単語の集合を**重み付き単語リスト**と呼ぶ．

最終検索用質問は以下のステップにより生成される．

1. GCSを用いた良い重み付き単語リスト(GWS)の生成

1.1 重みの計算：GCSの要素である各クラスタについて，KAを用いて重み付き単語リストを生成する．具体的には，(KA-1)により文書-語間の関連度をマトリックス値として得る．その際，語の重みメジャーとしてLtLnu法を用い，正規化方法として語ベクトル2乗和=1を用いた．このステップでは，GCSの要素であるクラスタ毎に1個の重み付き単語リストが生成される．

1.2 単語の選択：生成された重み付き単語リスト毎に，重みが大きいn7個の単語のみを選び，その他の単語をその重み付き単語リストから取り除く．

1.3 重みの平均：すべての重み付き単語リストに関して，単語毎に重みの和を求め，その単語が現れる重み付き単語リストの数で割ることで，重みを平均する．得られた重み付き単語リストがGWSである．

2. BCSを用いた悪い重み付き単語リスト(BWS)の生成：ステップ1と同様にBWSを生成する．ただし，ステップ1における「良い」，GCS，GWS，n7は，それぞれ，「悪い」，BCS，BWS，n8に置き換える．また，ステップ1.3では重みを平均する際に単語が現れる重み付き単語リストの数で割ったが，ここでは単語が現れる重み付き単語リストの数の2乗で割る．

3. GWSとBWSの混合：最初に，GWSの要素である各単語について，その重みをn9倍する．次に，GWSの要素である各単語について，もしその単語がBWSにも含まれていればBWSでの重みをn10倍した数でGWSでの重み(n9倍されている)を割る．このようにして得られた重みを用いて，1個の重み付き単語リストを生成する．この重み付き単語リストを中間質問(IQ)と呼ぶ．IQにはBWSのみに現れる単語は含まれない．

4. IQとQPSの混合

4.1 単語の分類：IQとQPSを，1)IQのみに現れる重み付き単語リスト(OnlyIQ)，2)QPSのみに現れる単語に関する重み付き単語リスト(OnlyQPS)，3)IQとQPSの両方に現れる単語に関する重み付き単語リスト(Both)，の3個の重み付き単語リストに分ける．ただし，Bothに含まれる単語の重みは未定義としておく．

4.2 Bothの重みの計算：Bothの要素である各単語について，IQにおける重みをn11倍しQPSの重みと足し，Bothでの重みとする．

4.3 重みの調整：OnlyIQの要素である各単語について，その重みをn12倍する．OnlyQPSの要素である各単語について，その重みをn13倍する．Bothの要素である各単語について，その重み(4.2で計算した)をn14倍す

る．このようにして，IQ と QPS に含まれるすべての単語から構成される新しい重み付き単語リストが生成される．

5. 単語の選択：得られた重み付き単語リストについて，重みが大きい n15 個以下の単語のみを残して，その他の単語を取り除く．得られた重み付き単語リストが最終検索用質問である．

3.3 パラメータのチューニング

第 3.2 節で述べたように TREC のデータ用に特化した文書クラスタ QE では，15 個のパラメータ(n1, n2, ..., n15)が存在する．これらのパラメータの値は，TREC から既に提供されている過去の質問と正解を用いて決定した．

このパラメータ・チューニング作業のうち，重みの調整（第 3.2.4 小節のステップ 4.3）では，Both への乗数(n14)は他のパラメータ(n12 と n13)より大きな値にすると，最終検索結果の上位に正解文書を押し上げる効果があることがわかった．

3.4 TREC における総合評価

TREC7 における結果は 38 団体中 13 位，85 の検索結果中 27 位となった．文書クラスタ QE は従来 QE より効果があった．TREC 用システムは過去の質問と正解を用いてチューニングした．チューニングには用いていない TREC7 の質問に対しても従来 QE よりも成績が良かったことより，文書クラスタ QE は検索質問の傾向に関して安定した手法であると言える．これは第 1 章で挙げた問題(P-3)の解決手段となっていることを示している．

4. 考察

4.1 文書クラスタ QE の効果

文書クラスタ QE には手法の選択肢が多く，また各手法は多くのパラメータを持っているので，連携する検索エンジンや対象文書の変更にフレキシブルに対応できる．関連要素技術と今回開発した技術の各機能を perl スクリプト，Makefile，KA-batch スクリプトなどにより統合することで，第 1 章で挙げた変更に対応した開発工数の問題(P-2)を解決している．

文書クラスタ QE による文書選択では，選択する文書が少数の上位文書ばかりに偏らないよう自動的に制御し，下位にあっても正解文書であれば選択する可能性を高めることができる．これは第 1 章で挙げた問題(P-1) を解決している．

さらに，選択した文書からの重み付き単語リスト生成で用いる関連度は検索対象文書から生成する．このため検索対象文書の傾向の変化に対しても安定した性能を発揮する．これは第 1 章で挙げた問題(P-2) を解決している．

4.2 今後の課題

文書クラスタ QE では他の統計的情報検索技術と同様に検索対象の文書の性質や検索質問の傾向（検索目的）が変更されるとチューニングが必要になる．また，他の QE と同様に検索エンジンが変更されるとチューニングが必要になる．文書クラスタ QE はこれらのチューニングを行うことができる．

しかしながら，そのチューニングにおけるノウハウは体系化されておらず，当事者の暗黙知として存在しているのみである．今後の課題としては二つのアプローチが考えられる： 1）チューニング・ノウハウを体系化し，検索エンジンチューニングエキスパートシステムを構築する，2）チューニングの必要のない，安定した性能を発揮する手法を開発する．文書クラスタ QE を以上のように発展させることを考えている．

5. おわりに

我々は文書クラスタリングエンジンを用いて，検索質問展開に文書クラスタリング技術を適用した．文書クラスタ QE は，予備検索の結果上位 M 件から単に上位 N 件を用いるのではなく，N 個の最良の文書クラスタを用いて検索質問展開を行う．TREC に参加したところ，文書クラスタ QE は従来 QE よりも良い結果を生成した．その原因として 1）文書クラスタ QE がフレキシブルであること，2）クラスタリングを用いた QE は質問群の差に対して安定していたこと，が考えられる．今後はチューニングの重要さに注目して研究を進めたい．

謝辞 KA, KAcluster の改造をこころよく引き受けていただき，また貴重な意見と助言をいただいた（株）富士通研究所の渡部勇氏に感謝します．検索エンジン Terass の研究開発者であり，連携に関する作業，貴重な意見と助言で協力いただいた同研究所の難波功氏，井形伸之氏に感謝します．また研究の機会を与えてくださった同研究所の松井くにお氏，松尾和洋氏に感謝します．

参考文献

[1] David Ellis：情報検索論 認知的アプローチへの展望，細野公男 監訳，丸善株式会社，1994

[2] Peter Ingwersen：情報検索研究 認知的アプローチ，藤原鎮男 監訳，トッパン，1995

[3] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford: Okapi at TREC-3, TREC3 Proceedings, 1994

[4] http://trec.nist.gov/

[5] I. Namba, N. Igata, H. Horai, K. Nitta, and K. Matsui: Fujitsu Laboratories TREC7 Report, TREC7 Proceedings, 1998

[6] 渡部勇：発散的思考支援システム「Keyword Associator」第二版，計測自動制御学会 第 15 回システム工学部会研究会資料，pp.9-16，1994

[7] 渡部勇，三末和男，新田清，杉山公造：ハイブリッド発想支援システム「HIPS」，計測自動制御学会 第 17 回システム工学部会研究会資料，95-PG-0001，pp.77-84，1995

[8] 松井くにお，難波功，井形伸之：高速テキスト検索エンジン，情報処理学会研究報告，Vol.97，No.49，ISSN 0919-6072 97-DD-7，デジタル・ドキュメント 7-3，pp.15-21，1997

[9] C. Buckley, A. Singhal, M. Mitra(, and G. Salton): New Retrieval Approaches Using SMART: TREC 4, TREC4 Proceedings, 1995

[10] Gerard Salton: Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer, p.318, Addison Wesley, 1989

[11] 鷲尾泰俊，大橋靖雄：多次元データの解析，岩波書店，p.234，1989

[12] William B. Frakes and Ricardo Baeza-Yates eds., Information Retrieval -- Data Structures & Algorithms, Prentice Hall, 1992