社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－159　(7)
2004／1／13

クラスタ指向インデクシングに関する一検討

相澤 彰子

国立情報学研究所

(akiko@nii.ac.jp)

本稿では、互いに類似するテキスト文書をグループ化し、資源として活用する情報検索システムの枠組みや構成法について検討する。提案手法ではまず、単語分布の類似度や単語列の一致度に基づき文書をクラスタリングして、共通する特徴語やフレーズの情報とあわせて文書クラスタを生成する。次に、文書クラスタを仮想的な１つの文書とみなして拡張インデックスを作成する。検索時には文書を単位とする通常の検索に加えて拡張インデックスによる検索を行い、両者を統合した結果を表示する。異なる観点に基づく文書クラスタリングの結果を、独立した拡張インデックスという形で付加的に定義することで、柔軟性のある検索の実現が可能であることを簡単な実例により示す。

An Approach to Cluster-based Indexing

Akiko AIZAWA

National Institute of Informatics

This paper introduces a framework and implementation of an information retrieval system that utilizes clusters of similar documents. The proposed method first generates document clusters together with their representative terms and phrases based on the term distribution or term sequence match. Next, considering each document cluster as a single virtual document, an extended index is created. Upon a query submission, the system uses both the original and the extended indices and returns the integrated result. An example is shown where indices generated based on different viewpoints are used to enhance the flexibility of the retrieval system.

1　はじめに

本稿では、テキスト情報検索における「クラスタ指向インデクシング」(cluster-based indexing) のアプローチに関する検討結果を報告する。本稿で提案するクラスタ指向インデクシングの基本的な考え方は、あらかじめ強い結びつきを持つ文書グループを検出し、グループへの寄与の度合いが高い出現語とともに比較的小さなサイズの文書クラスタを構成、これを情報資源として蓄積して検索に役立てようというものである [1]。

テキスト文書の検索において、文書間の類似が有力な手がかりとなることは、過去に多くの文献において指摘されてきた [2]。たとえば、Rijsbergen（1979年）は “the clustering hypothesis” として「互いに結びつき（association）が強い文書は同じ検索要求に対して高い関連度を示す傾向がある」と述べている [3] 。ここで、この場合の文書どうしの「結びつき」とは典型的には文書中の出現語パターン（すなわち文書ベクトル）の類似性に基づくものであり、検索語に対する文書の適合性判定と同様に計算される。すなわち類似文書のクラスタリングと適合文書集合の識別は互いに密接に関係する問題である。

このことから伝統的に情報検索における文書クラスタリングは、しばしば検索語ベクトルの生成・修正の問題と対をなすものとして研究されてきた。その中で今日もっとも広く知られているのは、Rocchioらによる適合フィードバック（relevance feedback）

である [4]。適合フィードバックは、利用者が検索結果の文書を適合／不適合にグループ分けした結果に基づき、検索語ベクトルを修正する方法である。適合フィードバックはその後の多くの情報検索に関する研究の中で参照されており、近年では Takano ら（2001 年）が、検索要求に合致する語と適合する文書を利用者に相補的に選択させながら検索を進める高速な汎用連想計算エンジンを実現している [5]。

上記の手法が個々の文書を利用者に提示するのに対して、文書クラスタを直接利用者に提示する方法も提案されている。Cutting ら（1992 年）の Scatter/Gather は、検索結果文書の自動クラスタリングと、利用者の手による適合クラスタの選択を交互に繰り返しながら検索を進めるブラウジングのためのインタフェースである [6]。また、Zamir&Etzioni （1998 年）は、接尾辞木構造を利用して検索結果のクラスタリングを高速に行う手法を提案し、Web 文書のブラウジングにおける有効性を示している [7]。江口ら（1999 年）は、適応的文書クラスタリングを適合フィードバックと組み合わせ、漸次的に検索語拡張を行う方法を提案している [8]。

一方、利用者が介在することなく自動的に検索語ベクトルの修正を行う方法も提案されている。Attar&Frankel（1977 年）の局所フィードバック（local feedback）は、上位にランキングされた検索結果文書中での共起関係に基づき語どうしの類似性を計算し、これを利用して検索語ベクトルの修正を行う方法である [9]。また、文書集合全体における共起関係から語どうしの類似度を計算する大域的方法として、Qiu&Frei （1993 年）の「概念（concept）」に基づく検索語拡張がある [11]。さらに、Xu&Croft （1996 年）の自動局所分析（automatic local analysis）では、文書集合全体から大域的な「概念」どうしの類似度を計算するとともに、単語のかわりに「概念」を用いた局所フィードバックで検索語ベクトルを修正する方法を用いている [10]。一方、金沢ら（1999 年）は大域的な類語関係を用いて、検索語ベクトルではなく、文書ベクトルを修正するモデルを提案している [12]。

ここで、これらの従来手法の多くに共通しているのは、まず、文書や語のグループが検索の過程で適応的に生成される点である。検索語拡張用のシソーラスを検索対象文書全体から自動抽出する場合を除き、クラスタを明示的にシステムに蓄積することは想定していない。また、従来手法における検索の主要な目的は個別の文書の順位付けであり、検索結果文書のクラスタリングは、利用者の適合判定を支援するためのブラウジング手段として用いられる場合が多い。比較的少数のクラスタを一瞥性よく提示することが前提となるため、クラスタ間での適合度の順位はあまり問題にならない。

これに対して本稿では、あらかじめ抽出した文書クラスタを仮想的な文書として扱い、拡張インデックスとして蓄積して検索に利用するシステムの構成を検討する。文書クラスタをインデックスすることの利点として次をあげることができる。

- 類似文書が複数個存在する場合に、これらをまとめることによって、個々の文書を単位とする場合よりもグループとして上位に順位付けするとができる。
- グループ化により適合度の順位が上がる効果に注目すると、特定の分野や話題に限定して文書クラスタを生成して、検索に付加的に分野指向性を持たせることができる。

以下、２節で文書クラスタの定義とそれを用いた検索システムの構成について述べる。３節では、単語分布と単語列一致に注目した２つの文書クラスタ生成法について説明し、次に４節でインデックスの重み計算と統合法を示す。最後に５節で簡単な実行例を紹介し、６節でまとめを述べる。

2　システム概要

2.1　文書クラスタの定義

検索対象となる文書の集合を D、D に含まれる語の集合を T として、「文書クラスタ」を、文書集合 S_D（$\in D$）とこれを代表する語集合 S_T（$\in T$）の組み合わせで定義する。文書クラスタの記述形式の例を図１に示す。クラスタ毎に S_T、S_D に加えて、画面表示用のテキスト情報を定義してある。

2.2　システムの構成

検索システムにおける処理の流れを図２と図３に示す。基本的な手順は、(1) 文書クラスタの生成、(2) インデックスの作成、(3) 検索とインデックス統合、の３つである。以下、各々について簡単に述べる。

(1) 文書クラスタの生成

本稿では文書クラスタとして、意味的なまとまりのある比較的小さな単位のグループを想定している。

```
<CLUSTER>
<SUMMARY>文書クラスタ表示用テキスト</SUMMARY>
<DOCSIZE>m</DOCSIZE>
<DOCLIST>
<DOC><ID>文書 ID1</ID><TEXT>文書タイトル 1</TEXT></DOC>
<DOC><ID>文書 ID2</ID><TEXT>文書タイトル 2</TEXT></DOC>
...
<DOC><ID>文書 IDm</ID><TEXT>文書タイトル m</TEXT></DOC>
<DOCLIST>
<TERMSIZE>n</TERMSIZE>
<TERMLIST>
<DOC><ID>語 ID1</ID><TEXT>語 1</TEXT></DOC>
<DOC><ID>語 ID2</ID><TEXT>語 2</TEXT></DOC>
...
<DOC><ID>語 IDn</ID><TEXT>語 n</TEXT></DOC>
</TERMLIST>
</CLUSTER>
```

図 1: 文書クラスタの記述形式例

このようなクラスタを生成する方法として、たとえばDhillonら（2001年）は、排他的に文書をクラスタリングして各々から抽出した代表ベクトルを「概念」とする概念分解（concept decompositions）法を提案しており [13]、文書 - 語行列の主成分分析に基づくLSI（Latent Semantic Indexing）[14] にかわる手法として注目される。また近年では、Slonim&Tishby（2000年）の Information Bottleneck 法、Dhillonら（2003年）の情報理論的 co-clustering 等、文書と語の相互情報量に基づく相互クラスタリング法も提案されている。本システムでは、次節で述べるように、情報量に基づく語と文書の相互的なクラスタリング [1]、および、単語列一致に基づく文書クラスタリング [15] を適用する。これらは、大規模かつ疎なデータ構造への対応を意識した局所的な文書クラスタリング法である。

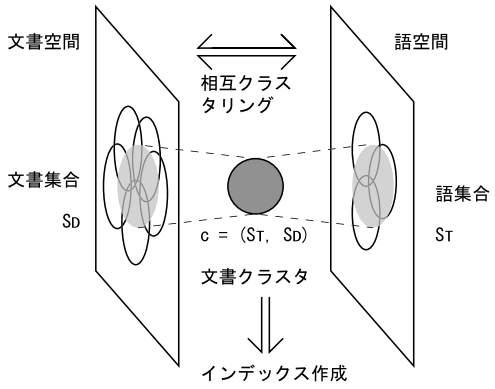

図 2: 文書クラスタの生成

(2) インデックスの作成

文書を単位とする通常の検索用インデックス（以下「基本インデックス」）と、文書クラスタを単位とするインデックス（以下「拡張インデックス」）をそれぞれ生成する。拡張インデックスは、文書クラスタ内の代表的な語や語句を、クラスタを構成する文書に対応付けるものである。ここで、現実システム上には多数の文書が存在するため、資源の時間的・空間的な制約のもとで、網羅的な文書クラスタリングを行うことは現実的ではない。このため、語彙や文書カテゴリを限定した上で拡張インデックスを作成することが必要になる。

(3) 検索とインデックス統合

前述のように、本稿における拡張インデックスの目的は、類似文書をグループ化することで、より高い適合度順位を与えようというものである。基本インデックスと拡張インデックスを独立に構成することで、利用者による明示的なインデックスの選択が可能になり、検索に付加的に分野指向性を持たせることができる（図 3）。拡張インデックスに対する語の重み付けとインデックス統合については 4 節で述べる。

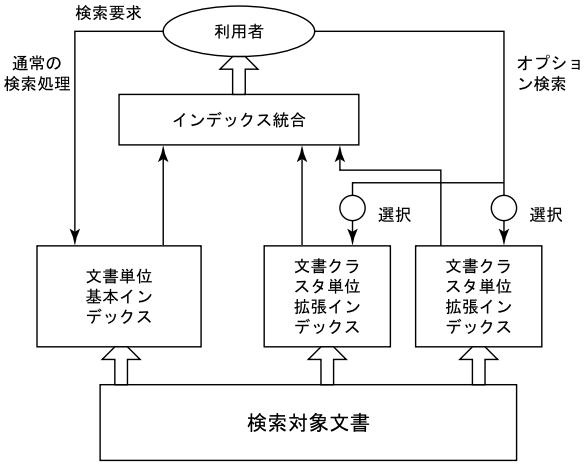

図 3: 検索とインデックス統合

3 文書クラスタの抽出

文書クラスタを抽出するためのアプローチとして、(i) 単語分布の類似度に基づくクラスタリング、および、(ii) 単語列の一致に基づくクラスタリング、の 2 つについて述べる。

3.1 単語分布一致に基づく方法

単語分布の一致に基づく方法では、文書を順序を持たない語の集合とみなして、単語分布が部分的に

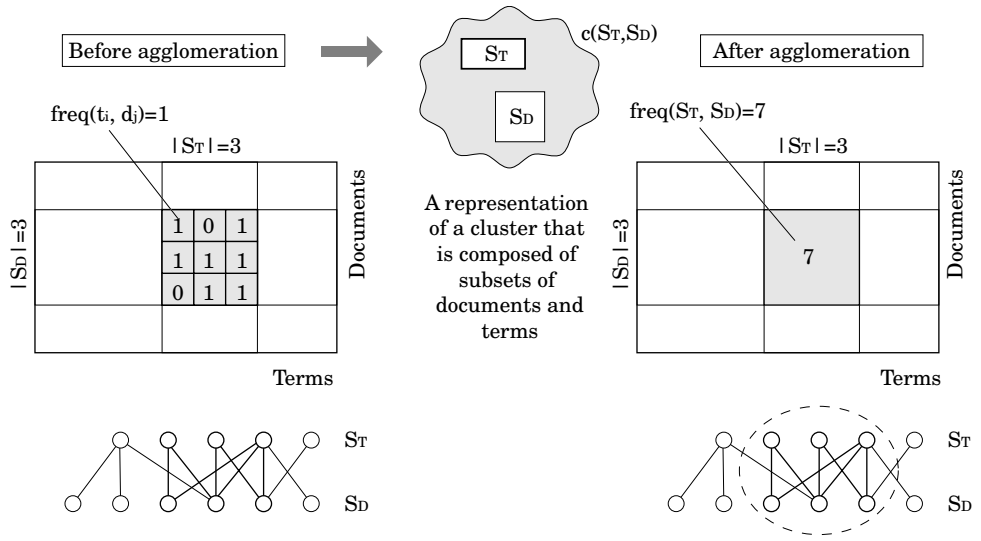

図 4: 単語分布に基づく文書クラスタリング法 [1]

類似する文書グループとそれに対応する単語グループを抽出する。具体的には、自動適合フィードバックと類似の仕組みを用いて、以下のように文書クラスタを抽出する [1]。

(1) ランダムに語を選択する。
(2) 選択した語を検索語として、それを含む文書を取り出す（S_D^* とする）。
(3) 抽出した文書に含まれる語を取り出す（S_T^* とする）。
(4) S_D^* および S_T^* に相互情報量に基づく基準を適用して、互いに関連が強い集合 (S_T, S_D) をとりだし文書クラスタとする。

手順 (4) では、図 4 に示すように、語と文書の共起確率行列の上で、S_T^* および S_D^* によって特定される領域に注目し、(S_T^*, S_D^*) 内の個々の語と文書の共起 (t_i, d_j)（$t_i \in S_T^*$, $d_j \in S_D^*$）をそれぞれ独立の事象とみなす場合と、(S_T^*, S_D^*) 全体をまとめて 1 つの事象とみなす場合との相互情報量の差分を計算しながら逐次的に語や文書を集合から取り除いて行く。この場合の文書クラスタの粒度は、頻度が小さいことによる確率の推定誤差と、個別の事象をまとめることによる情報の損失のトレードオフで定まる。確率誤差による損失は、確率的言語モデルにおけるディスカウンティングを使って計算している。

結果として抽出される S_T, S_D の各要素には、文書クラスタに対する寄与の度合いにしたがって、和が 1 になるように正規化した値で重みが付けられている。この重みにしたがって S_T の中から上位 M（$M = 10$）語を文書クラスタの内容を要約する表示用のキーワードとして抽出する。また検索用のキーワードには、S_T の単語すべてを用いる。なお、手順 (1) で選択する語を特定の分野に限定することで、偏りのあるクラスタを生成することができる。

3.2 単語列一致に基づく方法

単語列一致に基づく方法では、比較的長い単位で語句を共有する文書グループを抽出する。現実世界のテキスト集合には、このような長い単位での字句の一致がみられる文書が数多く存在するが、実際にテキストを分析した結果によると、これらの文書どうしは高い確率で参照関係にあり、相互に話題が共通している。また反復して利用される語句は往々にして有用な情報を含んでおり、たとえば新聞記事の場合には重要な説明文や組織名や人物の肩書きのように、要約の手がかりとしても有効に使うことができる。

これに基づき単語列一致に基づく方法では、まず単語を単位とする接尾辞木を構成して、情報量による一致度が閾値以上の単語列を数えあげ、これらの単語列を共有する文書グループを抽出する。ここで閾値は、あらかじめカテゴリが付与された新聞記事等のテキストを実際に分析することによって定めた [15]。この方法は基本的に接尾辞木を作成するコストで処理が行えるため大規模な文書に適している。

文書クラスタの抽出後は、各クラスタごとにもっと

も一致度が高い語句を取り出し、先頭から N （$N = 50$）語をクラスタ表示用のテキストとして抽出する。また、検索用のキーワードには S_D に共通する単語すべてを用いる。なお、類似度計算に分野指向性を持たせるために、実装上はあらかじめ登録した語彙に限定した一致度の計算を可能にしている。

4 インデクスの生成と検索方法

4.1 インデックス統合のための重み付け

異なるインデックス間で検索結果を統合するためには、両者に共通する文書ベクトルの重み付け基準が必要である。ここでは簡単のため、*tf-idf* を用いる場合を想定し、拡張インデックスの取り扱いについて検討する。

まず情報量的なモデルにしたがえば、語と文書の間の相互情報量は以下のようになる [17][1][16] 。

$$I(T,D) = \sum_{t_i \in T} \sum_{d_j \in D} P(t_i, d_j) log \frac{P(t_i, d_j)}{P(t_i)P(d_j)} \quad (1)$$

ただし T は語集合、D は文書集合、$P(\cdot)$ は語や文書を事象とみなした場合の生起確率である。いま、文書サイズや文書内での単語の分布に大きな偏りがないとすると、N_i を語 t_i を含む文書の数、N を文書の総数として、$P(d_j|w_i) \approx \frac{1}{N_i}$, $P(d_j) \approx \frac{1}{N}$ となる。このとき、*tf-idf* における *idf* は式 (1) の対数部分に対応するとみなせる。具体的には、基本インデックスの語 w_i の *idf* 値は以下となる。

$$idf_b(w_i) = log \frac{P(d_j|w_i)}{P(d_j)} = log \frac{N}{N_i} \quad (2)$$

次に、拡張インデックスで文書クラスタ (S_T, S_D) 中の w_i に対する *idf* 値を同様に計算すると、$|S_D|$ をクラスタ中の文書数として $P(S_D|w_i) \approx \frac{|S_D|}{N_i}$, $P(S_D) \approx \frac{|S_D|}{N}$ を用いて次式のようになり、

$$idf_e(w_i) = log \frac{P(S_D|w_i)}{P(S_D)} = log \frac{\frac{|S_D|}{N_i}}{\frac{|S_D|}{N}} = log \frac{N}{N_i} \quad (3)$$

基本インデックスにおける idf 値と等しい。すなわち、拡張インデックスを作成する場合には、単純に索引語を文書毎の出現回数の総和だけ含む新たな仮想的文書を想定すればよい。語や文書に関する統計情報は基本インデックスと共通した値を用いればよく、新たに計算しなおす必要はない。

4.2 文書クラスタ統合

文書クラスタ間に重複がある場合に、上記で得られた文書クラスタをスコア順にそのまま提示すると、同じ画面内に類似したクラスタが何度も出現する場合がある。そこでインデックス統合時に簡単な重複チェックを行うものとして、文書クラスタ中に含まれる既表示の文書の割合が r（$r = 50\%$）以上のものは順位リストから削除する。なお、各文書は「基本インデックス」内で単独でも索引付けされているため、この削除操作によって表示順が変わる場合でも、表示からもれることはない。

最後に画面表示の際には、文書クラスタの作成時に同時に抽出した表示用テキストを利用者に提示する。

5 実行例

現在、全文検索エンジン namazu[18] を利用して、プロトタイプシステムの作成を進めている。日本語の大規模テストコレクションである NTCIR1 （約 30 万文書）を用いた検索の例を図 5 および図 6 に示す。この例では、各学会で発表された文献に付与された著者キーワードから分野別の用語辞書を作成して、文書クラスタの生成に利用している。図中では、単語分布に基づく方法で得られた文書クラスタを拡張インデックスとして、「学習システム」という検索語に対する結果を示している。同じ検索語を入力する場合でも、「計測自動制御学会」に注目した図 5 と、「日本建築学会」に注目した図 6 では文書間で異なるランキングが得られることがわかる。

なお、２つの文書クラスタ生成法に関する数量的評価については、それぞれ文献 [1][15] に示している。また検索システムの数量的な評価については現在検討中であり今後の課題となっている。

6 おわりに

本稿では、文書クラスタを拡張インデックスとして蓄積／利用する検索システムの構成法に焦点を当て、その概要を述べた。システム実現の要となる重要な技術として、文書クラスタの生成および要約テキストの抽出の２つをあげることができる。文書クラスタ生成の手法は、本稿で述べた方法に限定されるものではなく、話題抽出やイベント追跡等における手法が適用可能である。また、適合フィードバックやソーシャルフィルタリングの結果を蓄積したり、Web 等のリンク解析を用いることも考えられる。表

図 5: NTCIR による実行例 (1)：計測自動制御学会

図 6: NTCIR による実行例 (2)：日本建築学会

示のための文書クラスタの要約テキストとして、本稿では単純に重要度の高い重要語や重要語句を用いたが、既存のテキスト要約手法の適用を含め今後の検討が必要である。

参考文献

[1] Akiko Aizawa 2002. A Method of Cluster-Based Indexing of Textual Data. Proc. of COLING 2002, 1-7. 2002.

[2] R. K. Belew Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW Cambridge University Press. 2000.

[3] C. J. van Rijsbergen. Information Retrieval Butterworth 1979.

[4] R. Baeza-Yates and B. Riberio-Neto. Modern Information Retrieval Addison-Wesley. 1999.

[5] A. Takano, Y. Niwa, S. Nishioka, T. Hisamitsu, M. Iwayama, and O. Imaichi. Associative Information Access using DualNAVI. Proc. of NLPRS 2001, pp. 771-772. 2001.

[6] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. Proc. of ACM SIGIR '92, pp. 318-329. 1996.

[7] Oren Zamir and Oren Etzioni. Web Document Clustering: A Feasibility Demonstration. Proc. of SIGIR '98, 46–54. 1998.

[8] 江口、伊藤、隈元、金田. 漸次的に拡張されたクエリを用いた適応的文書クラスタリング法 電子情報通信学会論文誌, D-I, Vol. J82-D-I, No. 1, pp. 140-149. 1999.

[9] R. Atter and A. S. Frankel. Local Feedback in Full-Text Retrieval Systems. Journal of the Association for Computing Machinery, Vol. 24, No. 3, pp. 397-417. 1977.

[10] Jinxi Xu and W. Bruce Croft. Query Expansion Using Local and Global Document Analysis. Proc. of ACM SIGIR '96, pp. 4-11. 1996.

[11] Y. Qiu and H. P. Frei. Concept Based Query Expansion. Proc. of ACM SIGIR '93, pp. 160-169. 1993.

[12] 金沢、高須、安達. 文書関連性を考慮した検索方式. 情報処理学会研究報告, データベース・システム, Vol. 98, No. 58, pp.165-172. 1998.

[13] I. S. Dhillon, S. Mallela, and D. S. Modha. Concept Decompositions for Large Sparse Text Data using Clustering. Machine Learning, Vol.42, No.1, pp. 143-175. 2001.

[14] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. Journal of American Society of Information Science, Vol. 41, pp. 391-407. 1990.

[15] Akiko Aizawa Analysis of Source Identified Text Corpora: Exploring the Statistics of Reused Text and the Authorship. Proc. of ACL 2003, pp. 383-390. 2003.

[16] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-Theoretic Co-clustering. Proc. of ACM SIGKDD 2003. pp. 89-98. 2003.

[17] N. Slonim and N. Tishby. Document Clustering Using Word Clusters via the Information Bottleneck Method. Proc. of SIGIR 2000, pp. 208-215. 2000.

[18] http://www.namazu.org/

[19] http://research.nii.ac.jp/ntcir/