目　次

1　はじめに

1.1　研究の背景 1

1.2　研究の目的 2

1.3　本論文の構成 2

2　関連研究

2.1　クラスタリング手法 3

2.2　クラスタを使った可視化 8

2.3　ユーザの視点を取り入れた再クラスタリング 10

2.4　重要語の判定による再クラスタリング 12

2.5　本システムとの関連 15

3　社会情報可視化システムの構築

3.1　システムの概要 16

3.2　システムの特徴 18

3.2.1 3 つのレベルでの文脈の可視化 18

3.2.2 適応クラスタリングによる個人的選好の反映 18

3.3　クラスタリング 19

3.3.1 索引語の抽出 19

3.3.2 索引語ベクトルの算出 20

3.3.3 類似度の計算 21

3.3.4 クラスタリング手法 21

3.3.5 適応クラスタリング手法 23

3.4　ユーザインタフェイス 24

3.4.1 キーワード選択画面 24
3.4.2 全体結果表示画面 25
3.4.3 特定記事の内容表示 25
3.4.4 特定クラスタの構造表示 26
3.4.5 適応クラスタリング手法 29
3.5 実行結果 31

4 適応クラスタリング
4.1 適応クラスタリングの概念 36
4.2 索引語頻度－逆文書頻度値を使った適応クラスタリング 39
4.3 χ^2 値を使った適応クラスタリング 40
4.4 文脈依存の度合いを使った適応クラスタリング 42
4.5 見なし共起を使った適応クラスタリング 44

5 適応クラスタリングの手法評価
5.1 データ 45
5.2 実験方法 47
5.3 結果と考察 48
5.3.1 索引語頻度－逆文書頻度値を使った方法の結果と考察 48
5.3.2 χ^2 値を使った方法の結果と考察 57
5.3.3 文脈依存の度合いを使った方法の結果と考察 66
5.3.4 見なし共起を使った方法の結果と考察 71
5.3.5 考察のまとめ 79

6 結論
6.1 本研究の成果 81
6.2 今後の課題 82
謝辞 83
参考文献
付録 プログラムリスト

図　目　次

２．１．１　クラスタリングの例 4
２．１．２　cat ate cheese，mouse ate cheese too，cat ate mouse too の Suffix Tree 7
２．２．１　Scatter/Gather システムのユーザインタフェイスのトップレベル画面の一部（[Hearst95]より引用） 8
２．２．２　Grouper システムのメイン結果ページ（[Zamier99]より引用） 9
２．３．１　Scatter と Gather の概念図（[Cutting92]より引用） 10
２．３．２　Grouper システムの検索質問の絞込み（[Zamier99]より引用） 11
２．４　新聞記事の構造（[福本 99]より一部変更の上、引用） 13
３．１　システムの概要 17
３．４．１　キーワード選択画面 24
３．４．２　全体結果表示画面 25
３．４．３　特定記事の内容表示画面 26
３．４．４．１　頂点間の理想距離 28
３．４．４．２　特定クラスタの構造表示画面 29
３．４．５．１　適応クラスタリングの選択画面 30
３．４．５．２　適応クラスタリングの結果画面 30
３．５．１　初期全体結果表示画面（環境－地球温暖化） 32
３．５．２　適応クラスタリング後の全体結果画面（環境－地球温暖化） 33
３．５．３　初期構造表示画面（環境－地球温暖化） 34
３．５．４　適応クラスタリング後の構造表示画面（環境－地球温暖化） 35
４．１　適応クラスタリングの概念図 37
４．４　新聞記事の構造（[福本 99]より一部変更の上、引用） 42
４．５　見なし共起を使った類似度の計算 44

５．３．１．１　核－核抑止力の TFIDF 法による適応クラスタリング……51
５．３．１．２　原子力－原発反対の TFIDF 法による適応クラスタリング……52
５．３．１．３　環境－地球温暖化の TFIDF 法による適応クラスタリング……53
５．３．１．４　福祉－政策の TFIDF 法による適応クラスタリング……54
５．３．１．５　通信－無線の TFIDF 法による適応クラスタリング……55
５．３．１．６　情報－情報公開の TFIDF 法による適応クラスタリング……56
５．３．２．１　核－核抑止力のχ²法による適応クラスタリング……60
５．３．２．２　原子力－原発反対のχ²法による適応クラスタリング……61
５．３．２．３　環境－地球温暖化のχ²法による適応クラスタリング……62
５．３．２．４　福祉－政策のχ²法による適応クラスタリング……63
５．３．２．５　通信－無線のχ²法による適応クラスタリング……64
５．３．２．６　情報－情報公開のχ²法による適応クラスタリング……65
５．３．３．１　核－核抑止力の文脈依存の度合いによる適応クラスタリング……69
５．３．３．２　原子力－原発反対の文脈依存の度合いによる適応クラスタリング……69
５．３．３．３　環境－地球温暖化の文脈依存の度合いによる適応クラスタリング……69
５．３．３．４　福祉－政策の文脈依存の度合いによる適応クラスタリング……70
５．３．３．５　通信－無線の文脈依存の度合いによる適応クラスタリング……70
５．３．３．６　情報－情報公開の文脈依存の度合いによる適応クラスタリング……70
５．３．４．１　核－核抑止力の見なし共起による適応クラスタリング……73
５．３．４．２　原子力－原発反対の見なし共起による適応クラスタリング……74
５．３．４．３　環境－地球温暖化の見なし共起による適応クラスタリング……75
５．３．４．４　福祉－政策の見なし共起による適応クラスタリング……76
５．３．４．５　通信－無線の見なし共起による適応クラスタリング……77
５．３．４．６　情報－情報公開の見なし共起による適応クラスタリング……78

表　目　次

５．１　　キーワードと正解セットの特性 46

５．３．１．１　索引語頻度－逆文書頻度値を使った方法で判別された重要索引語（上位１０語） 50

５．３．１．２　キーワードと正解セットの特性と索引語頻度－逆文書頻度法の結果 49

５．３．２．１　　χ^2値を使った方法で判別された重要索引語（上位１０語） 59

５．３．２．２　　キーワードと正解セットの特性とχ^2法の結果 59

５．３．３　文脈依存の度合いを使った方法で判別された重要索引語 68

５．３．４　キーワードと正解セットの特性と見なし共起法の結果 71

５．３．５．１　f 値が最大になる割合と重み 79

５．３．５．２　f 値が最高になる見なし共起の割合 79

第１章
はじめに

1.1　研究の背景

近年のインターネットの爆発的な普及は、我々を取り巻く情報環境を大きく変えつつある。特に WorldWideWeb（WWW）の普及によって我々が入手可能な電子化された情報の量は膨大なものとなった。このような変化にともない、我々の生活において情報を検索するという行為は身近で重要なものになりつつあり、その手段としてさまざまな情報検索システムが現れてきている。

現在最も有力な情報検索システムのひとつである Web 検索エンジンで、例えば「環境」を検索してみると、例えば www.google.co.jp で約 1,420,000 件、www.goo.ne.jp で約 2,400,000 件の関連 Web サイトが検索され、全てのサイトにアクセスする事は不可能である。これほど大量のサイトが検索される１つの理由は、「環境」という単語が多種多様な意味で使われているからであり、ユーザが検索したい意味の「環境」を表しているサイトに絞り込む必要がある。１つの方法として、さらに別のキーワードで絞り込む事が考えられる。しかし、少数の一般的なキーワードでは十分にサイトを絞り込むことは出来ず、逆に複数の特殊なキーワードで絞り込みすぎると全く関連サイトを得られないことになる。特にユーザが当該分野の門外漢である場合は、適切なキーワードの選択は難しい。別の方法としては、www.yahoo.co.jp のように予め Web サイトをカテゴリー分けしておき、検索しやすくしているサービスも存在する。しかし、例えば「環境」を www.yahoo.co.jp で検索してみると、172 種類のカテゴリーと 3059 件の Web サイトが検索される。検索される Web サイトの数が他の検索エンジンに比べて著しく少なくなっているとはいえ、一つ一つ見るには多すぎ

る量である。

従来の検索モデルでは答えが 1 つ存在し、それに関連する文書が順位付けられる。単純な問題ではこのようなモデルで十分であるが、社会的な懸案事項－例えば環境問題など－は、どのような経緯があったとか、どのように捉えられてきたかなど、情報を収集し、自分なりの視点を作り上げていくことが必要である。例えば「環境」が時代によってどのような意味で使われていて、どのように推移しているかを調べる場合を考えてみよう。この場合、「環境」の検索結果を時間順に並び替え、適当な時間間隔で抽出した代表的なサイトを読み、その変遷をまとめる必要がある。しかし、ユーザがこれを手作業で行うには、当該分野の専門知識と膨大な作業量が必要となる。

1.2 研究の目的

本研究の目的は、ユーザの視点を反映した社会情報の時間的変遷を可視化するシステムの構築である。

社会情報の時間的変遷の可視化には、クラスタリングによる情報の分類と、時間軸を持つ空間上に情報を配置する方法をとる。ユーザの視点の反映には、ユーザからのフィードバックをクラスタに反映させる適応クラスタリングの方法を使う。

1.3 本論文の構成

第 2 章では、本研究と関連する研究をまとめ、本研究の位置付けを明確にする。

第 3 章では、構築した社会情報可視化システムの概要、構築方法、操作方法、実行結果について述べる。

第 4 章では、ユーザの視点を反映させる方法としての、適応クラスタリングの様々な方法、特徴について述べる。

第 5 章では、第 4 章で説明した適応クラスタリングの性能を評価する実験について、その方法を述べ、結果を考察する。

第 6 章では、本研究の結論を述べる。

第 2 章
関連研究

本章では、最初にクラスタリング手法を簡単に説明し、そのクラスタリング結果の可視化方法に関する関連研究を紹介する。次にユーザの視点を取り入れた再クラスタリング時に必要となる重要索引語の判定方法に関する関連研究を説明し、最後に本研究との関連をまとめる。

2.1　クラスタリング手法

複数の文書を意味的に関連したグループに分類するのに用いられるクラスタリングの方法は、図 2.1.1 に示すように、階層型クラスタリング(hierarchical clustering)と非階層型クラスタリング(non-hierarchical clustering)に大きく分類することが出来る[徳永 99][岸田 98]。

階層型クラスタリングの結果は、文書を葉節点に持つ木構造(dendrogram)となり、木の中間節点はその節点の子節点からなるクラスタを表している。図 2.1.1(a)のd_i は個々の文書を、c_i はクラスタを表している。例えば、文書d_1 と d_2 はクラスタc_4 に属し、さらにクラスタc_2、c_1 にも属している。

非階層型クラスタリングの結果は、図 2.1.1(b)に示すように、文書がいくつかのグループに分類された平坦な構造になる。階層型クラスタリングで得た木構造において、重複した葉節点を持たない中間節点の集合を選ぶと、非階層型クラスタリングと同じ平坦なクラスタを得ることが出来る。例えば、図 2.1.1(a)の中間節点c_3, c_4, c_5 を選択すると、図 2.1.1(b)と同じクラスタとなる。

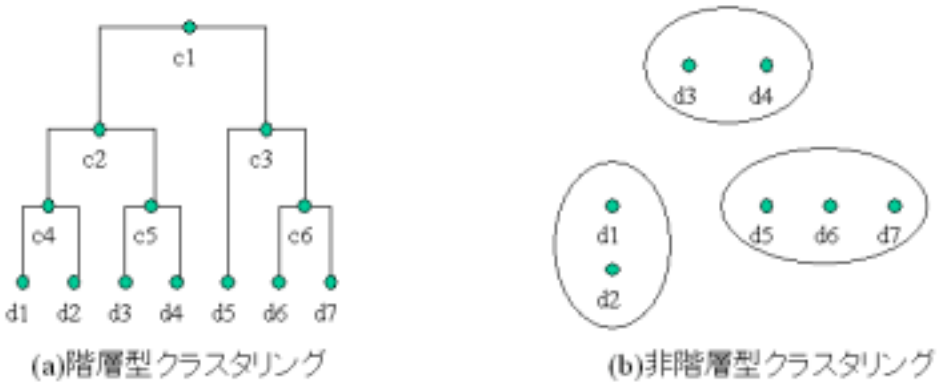


図 2.1.1　クラスタリングの例

階層型クラスタリングのアルゴリズムは、以下の手続きからなる。

1　各文書だけからなるクラスタを作る。
2　クラスタの数が１つになるまで以下を繰り返す。
　2.1　全てのクラスタの組の類似度を計算する。
　2.2　最も類似度の大きいクラスタの組を併合し、併合によって出来たクラスタと他のクラスタの類似度を計算する。

2.1 のクラスタの組の類似度の計算方法には、単連結法(single link method)、完全連結法(complete link method)、群平均法(group average method)、メディアン法(median method)、重心法(centroid method)、ウォード法(Ward method)などがある。代表的な方法の計算方法を以下に示す。

1 単連結法

クラスタ c_i に含まれるテキスト d_i とクラスタ c_j に含まれるテキスト d_j の類似度のうち、最も大きい値をクラスタ間の類似度とする。単連結法での類似度の計算方法を(2.1.1)に示す。

$$sim(c_i, c_j) = \max_{d_i \in c_i, d_j \in c_j} (sim(d_i, d_j)) \qquad (2.1.1)$$

2 完全連結法

クラスタ c_i に含まれるテキスト d_i とクラスタ c_j に含まれるテキスト d_j の類似度のうち、最も小さい値をクラスタ間の類似度とする。完全連結法での類似度の計算方法を(2.1.2)に示す。

$$sim(c_i, c_j) = \min_{d_i \in c_i, d_j \in c_j} (sim(d_i, d_j)) \tag{2.1.2}$$

3 群平均法

クラスタ c_i に含まれるテキスト d_i とクラスタ c_j に含まれるテキスト d_j の類似度の平均をクラスタ間の類似度とする。群平均法での類似度の計算方法を(2.1.3)に示す。ただし、N_i をクラスタ c_i に含まれるテキストの数、N_j をクラスタ c_j に含まれるテキストの数とする。

$$sim(c_i, c_j) = \frac{\sum_{d_i \in c_i} \sum_{d_j \in c_j} sim(d_i, d_j)}{N_i N_j} \tag{2.1.3}$$

単一連結法の場合、類似度の高い文書が一組でも存在すれば、他の文書間の類似度が低くても同じクラスタに併合されることになり、逆に完全連結法では、一組でも類似度の低い文書が存在すれば併合されない。この１つの類似度だけが大きな影響を与える欠点を補う方法が群平均法である。

非階層型クラスタリングのアルゴリズムの代表的なものに、単一パス法(single pass method)と再配置法(reallocation method)がある。

単一パス法の手順を以下に示す。

1. 最初の文書を最初のクラスタの代表とする
2. 次の文書と、その時点で存在するすべてのクラスタとの類似度を計算する。
3. 類似度にしたがって、その文書をいずれかのクラスタに割り当てる。もし、いずれのクラスタにも該当しなければ、それを代表とする新しいクラスタを生成する。

4. 最後の文書に達するまで２に戻る。

再配置法の手順を以下に示す。なお、再配置法はあらかじめ生成されるクラスタの数を決めることが出来る。

1. お互いに類似していない文書をクラスタの種として用意する。
2. 残りの文書を１つづつこれらの初期クラスタの中で最も類似しているものに加える。ここで、各クラスタの重心を計算しなおし、もう一度すべての文書に対して同じことを繰り返す。クラスタの重心とは、クラスタを構成している文書を表す索引語ベクトルの平均ベクトルである。
3. クラスタ間で文書の移動がなくなるまで２を繰り返す。

ここで、クラスタリング精度を保ちながら、クラスタリング速度を早くする方法をいくつか紹介する。

Scatter/Gather システムでは、非階層型クラスタリングアルゴリズムの再配置法をベースに様々な改良を加えている[Cutting 92]。まず、再配置法の最初の手順であるクラスタの種となる文書の発見方法として、Buckshot 法と Fractionation 法を考案している。Buckshot 法では文書集合をランダムにサンプリングすることにより文書数を減らし、このサブ集合に改良された群平均法のクラスタリングを適応することによりクラスタの種を見つけ初期クラスタとしている。また、Fractionation 法では、決められた数のグループに対して郡平均法のクラスタリングを連続して適応することにより、Buckshot 法より精度の高い種を発見している。次に、残りの文書を初期クラスタの中で最も類似しているものに加えているが、この類似度の計算にはクラスタを構成する文書の内、最も重心に近い文書だけを使い、クラスタのごみを拾わない工夫をしている。

Grouper システム[Zamir 99]では、Suffix Tree Clustring（STC）アルゴリズム[Zamir98]を使って効率的にクラスタリングを行っている。この STC アルゴリズムは、文書集合のサイズに比例する線形時間で計算が終了する、1 つの文書を重複したクラスタに属させる、クラスタリングにフレーズを使う、クラスタの数をあらかじめ決めないなどの特徴がある。

このアルゴリズムの論理的ステップを以下に挙げる。

1. 文書クリーニング
   数字や HTML タグなどを削除する。
2. ベースクラスタを決める
   Suffix Tree を用いてスコア付けされたフレーズの転置行列を作り、ベースとなるクラスタを決める。Suffix Tree とは文の後接語(Suffix)を木構造(trie)に展開したデータ構造である。 cat ate cheese，mouse ate cheese too，cat ate mouse too を Suffix Tree に展開した例を図 4.2.2 に示す。Suffix Tree のノードは円で表されており、1 つまたは複数の文を表す長方形のラベルが付いている。このラベルの最初の数字は文を表す番号、最後の数字は何番目の後接語かを表す。例えば、図 4.2.2 の一番左側のラベル 1.1 は１番目の文 cat ate cheese のを表し、cheese は cat ate に対しては１番目の後接語である事を示している。ノードは共通して現れるフレーズを表しているが、そのノードの集合が、１つの文書集合を表している。よって、各々のノードはその文書集合の中で共通して現れるフレーズということになり、これをベースクラスタとする。

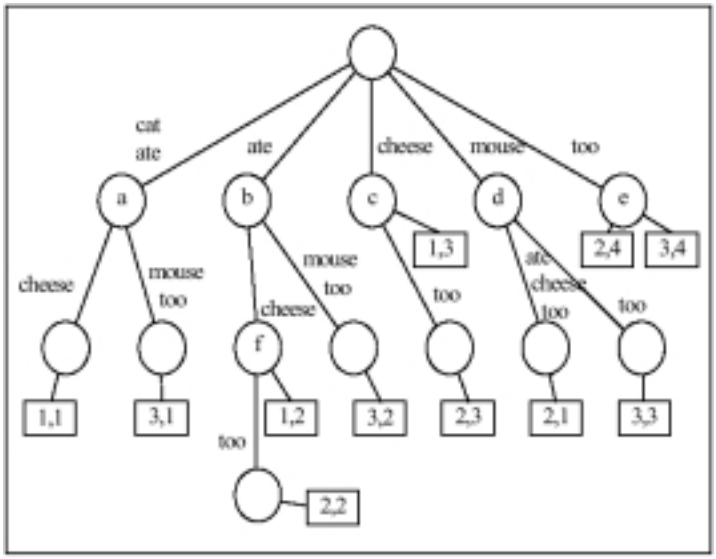


図 4.2.2　cat ate cheese，mouse ate cheese too，cat ate mouse too の Suffix Tree

3. ベースクラスタをまとめる
   フレーズの重複度を使ってベースクラスタをまとめる。このベースクラス

タのまとめには、単連結法と同様の方法を使っている。

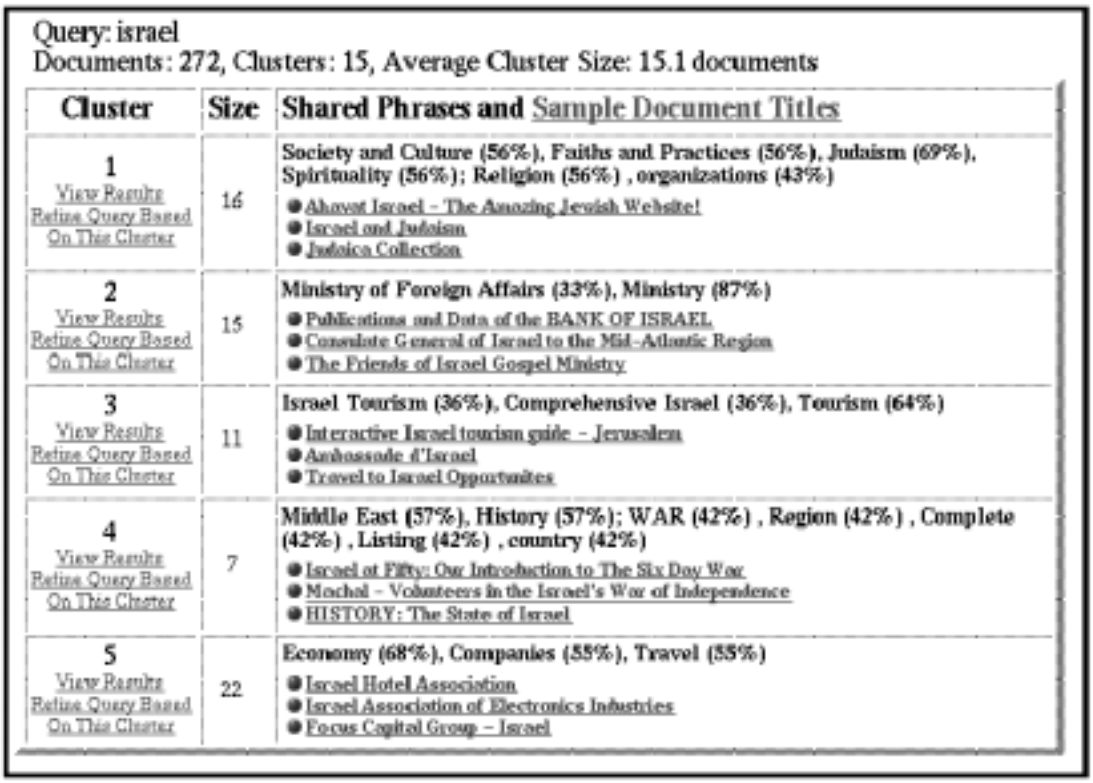
2.2　クラスタを使った可視化

本節では、クラスタリング結果を表形式で可視化する方法を紹介する。

Scatter/Gather システム[Hearst 95b][Cutting 92]では、クラスタリング結果は、各々のクラスタを要素とする表形式で表示される(図 2.2.1)。各々のクラスタは、文書数やそのクラスタに典型的な索引語やタイトルによって要約され、リスト表示されている。クラスタに典型的な索引語はそのクラスタ内で最も頻繁に出現した索引語が、典型的なタイトルはクラスタに重心に最も近い文書のタイトルが選ばれる。

図 2.2.1　Scatter/Gather システムのユーザインタフェイスの
トップレベル画面の一部（[Hearst 95]より引用）

Grouper システム[Zamir 99]では、クラスタリング結果は大きなテーブル上に表現される(図 2.2.2)。テーブルの 1 つの行が１つのクラスタを表している。各々のクラスタはクラスタに含まれる文書数、共通に現れるフレーズ、文書タイトルののサンプルで要約される。フレーズの横の括弧内の数字は、そのフレーズがクラスタ内に表れた全フレーズに対する割合を表す。



Query: israel
Documents: 272, Clusters: 15, Average Cluster Size: 15.1 documents

Cluster	Size	Shared Phrases and Sample Document Titles
1 View Results Refine Query Based On This Cluster	16	Society and Culture (56%), Faiths and Practices (56%), Judaism (69%), Spirituality (56%); Religion (56%) , organizations (43%) ● Ahavat Israel - The Amazing Jewish Website! ● Israel and Judaism ● Judaica Collection
2 View Results Refine Query Based On This Cluster	15	Ministry of Foreign Affairs (33%), Ministry (87%) ● Publications and Data of the BANK OF ISRAEL ● Consulate General of Israel to the Mid-Atlantic Region ● The Friends of Israel Gospel Ministry
3 View Results Refine Query Based On This Cluster	11	Israel Tourism (36%), Comprehensive Israel (36%), Tourism (64%) ● Interactive Israel tourism guide - Jerusalem ● Ambassade d'Israel ● Travel to Israel Opportunities
4 View Results Refine Query Based On This Cluster	7	Middle East (57%), History (57%); WAR (42%) , Region (42%) , Complete (42%) , Listing (42%) , country (42%) ● Israel at Fifty: Our Introduction to The Six Day War ● Machal - Volunteers in the Israel's War of Independence ● HISTORY: The State of Israel
5 View Results Refine Query Based On This Cluster	22	Economy (68%), Companies (55%), Travel (55%) ● Israel Hotel Association ● Israel Association of Electronics Industries ● Focus Capital Group - Israel

図 2.2.2　Grouper システムのメイン結果ページ([Zamir 99]より引用)

2.3　ユーザの視点を取り入れた再クラスタリング

クラスタリング結果にユーザーの視点を取り入れる方法の例として、Scatter/Gather システムと Grouper システムを説明する。まず Scatter/Gather システムでは、検索結果をクラスタリングすることにより散り散り（Scatter）にし、クラスタごとに短い要約をつける。ユーザはこの要約を読み、いくつかのクラスタを選択する。このユーザが選択したクラスタは集められ（Gather）サブ集合となる(図 2.3.1)。このサブ集合に対して再び Scatter/Gather を適用することにより、クラスタは小さく詳細になり、検索結果を絞る込むことになる。このシステムでは、このようにユーザ自身の視点でクラスタを集め、システムにフィードバックすることが出来る。

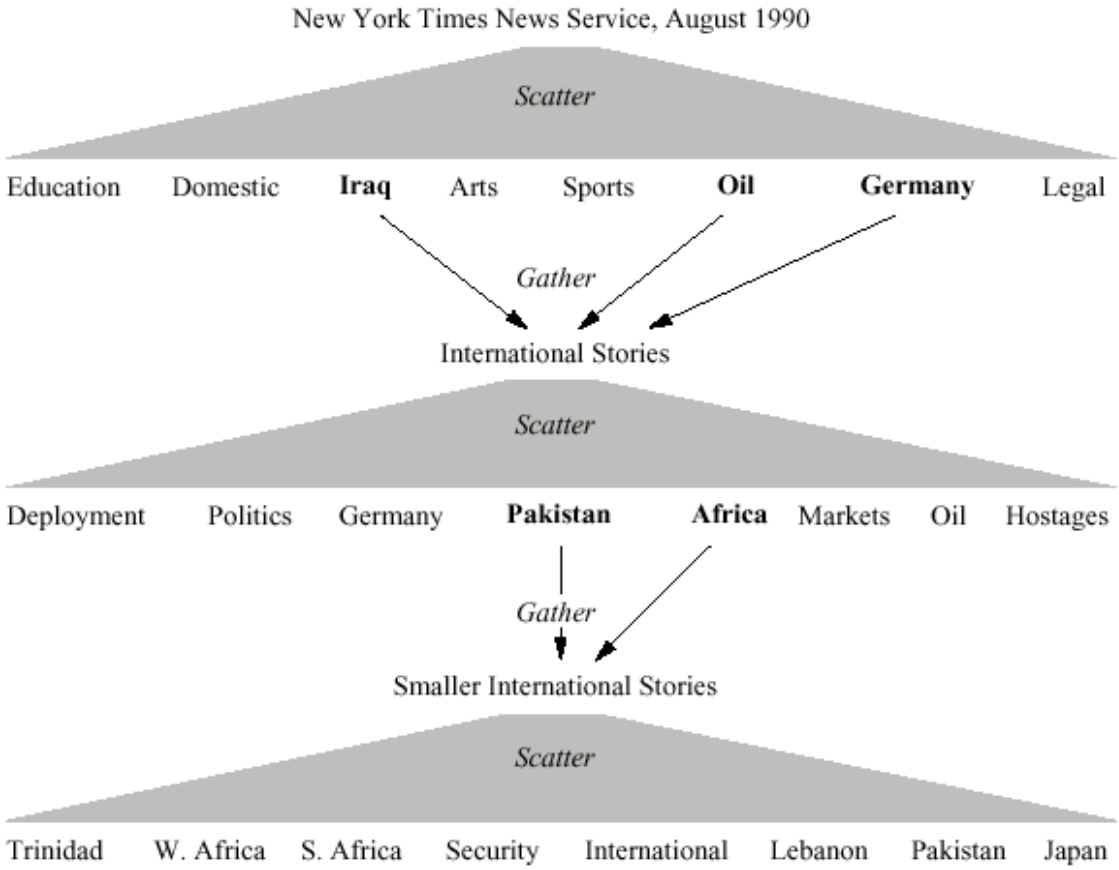


図 2.3.1　Scatter と Gather の概念図（[Cutting 92]より引用）

一方、Grouper システムでは、ユーザーの視点を取り入れる方法として、メイン結果ページ(図 2.2.2)のクラスタのうち 1 つを選ぶことにより、そのクラスタに特徴的な索引語やフレーズを検索語に加えることが出来る(図 2.3.2)。新たな検索語はオリジナルの検索語と論理積をとって検索され、検索結果を絞り込むことになる。

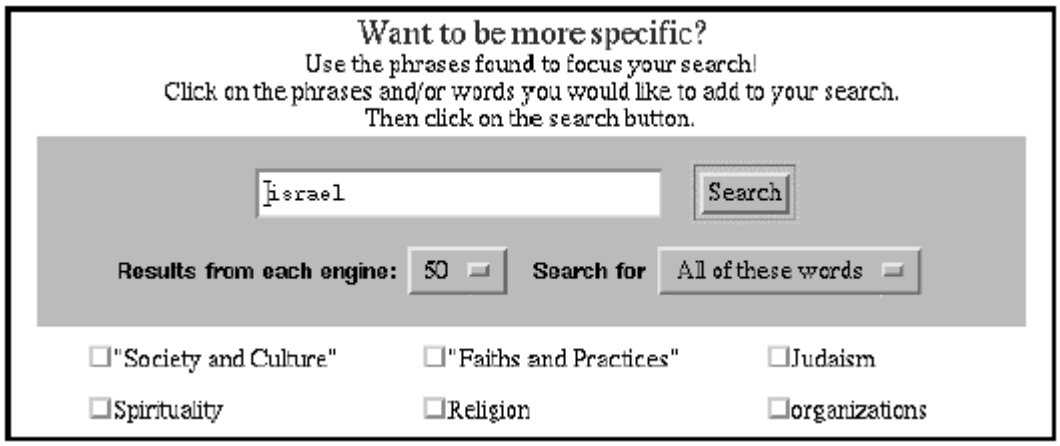


図 2.3.2　Grouper システムの検索質問の絞込み([Zamir 99]より引用)

2.4　重要語の判定による再クラスタリング

文書の再分類の手法として、クラスタリングの対象となる索引語のうち、重要なものを判別し、その重みを変えて再クラスタリングする方法がある[清田 98][福本 99]。この場合、どのような方法で重要な索引語を判定するかが重要となる。

最初に、χ^2値を使った重要語の抽出について説明する。一般にχ^2検定とは、2 つの要因の独立性を検定する方法である。ここでは、要因として記事内における索引語の出現頻度と、クラスタ内における索引語の出現頻度を採用する。もし、ある索引語が全記事中で全く均質に分布している場合は、2 つの要因は全く無連関の状態にあり、χ^2値は最小値０となる。逆に、ある検索語がある特定のクラスタ内にしか出現しない場合は、最大連関の状態にありχ^2値は最大値をとる。よって、χ^2値により、ある索引語の記事内とクラスタ内の出現頻度の偏りを計算することが出来る。このχ^2法を用いて重要なキーワードを抽出する研究が行われており[長尾 76]、χ^2法がキーワードの抽出に有効であることが確かめられている。また、重要漢字の自動抽出にχ^2法を用いる研究もされている[渡辺 94]。しかし、χ^2値をそのまま使う方法では、記事数の多いクラスタのχ^2値は大きく、逆に記事数の少ないクラスタのχ^2値は小さくなってしまう。この問題を解決するため[渡辺 94]はそれぞれのクラスタにおける出現頻度の理論度数からのずれに着目する方法を採用している。理想頻度とは、全記事に等確率でその索引語が出現した場合の出現頻度である。

検索語wがi（iは記事またはクラスタ）において特定の記事（またはクラスタ）jに依存する度合いを式(2.4.1)に示す。

$$(\chi^2)^i_{wj} = \begin{cases} \dfrac{(x_{wj} - m_{wj})^2}{m_{wj}} & if \quad x_{wj} > m_{wj} \\ 0 & otherwise \end{cases} \tag{2.4.1}$$

ここで、

$$m_{wj} = \frac{\sum_{j=1}^{n} x_{wj}}{\sum_{w=1}^{m}\sum_{j=1}^{n} x_{wj}} \times \sum_{w=1}^{m} x_{wj}$$

ただし、

i	記事、またはクラスタ
m	索引語の数
n	記事またはクラスタの数
x_{wj}	特定の記事またはクラスタ j における索引語 w の出現頻度
m_{wj}	特定の記事またはクラスタ j における索引語 w の理想頻度

次に、文脈依存の度合いを用いて重要語を抽出する方法を説明する[福本 99]。

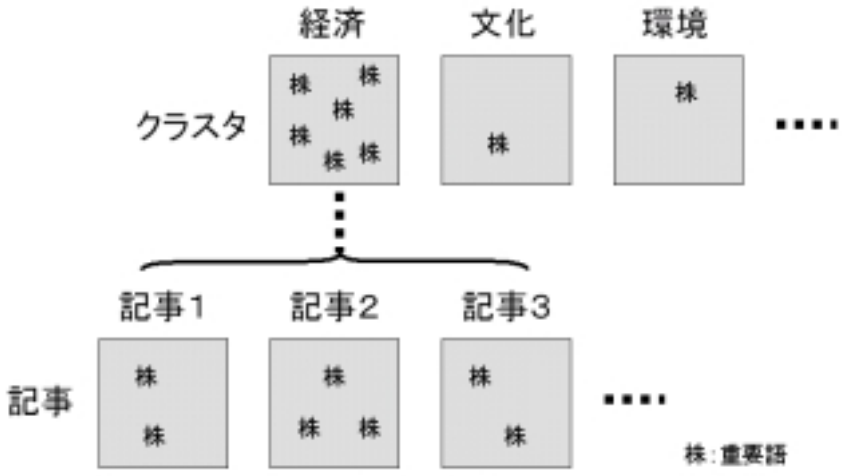


図 2.4　新聞記事の構造　（[福本 99]より一部変更の上、引用）

図 2.4 で示される新聞記事の構造において、文脈依存の度合いとは、ある索引語がこの図で示した特定のクラスタ、あるいは特定の記事とどのくらい深く関っているかという度合いの強さを示す。例えば、図 2.4 において、‘ 経済 ’ に関するクラスタにおける重要索引語を ‘ 株 ’ とすると、‘ 株 ’ は各記事にまたがり出現する。よって ‘ 株 ’ は各記事での分布の偏りが一般語と同様、小さく、特定の記事に依存する度合いは低い。次に ‘ 経済 ’ のクラスタについて考える。一般語は色々な文書中に均等に現れるため、各クラスタにおける分布の偏りと、記事における分布の偏りに差はない。一方、‘ 株’の ‘ 経済’での依存の度合いは ‘ 株 ’ が ‘ 経済 ’ という特定のクラスタに集中して出現するため、結果的に特定の記事に依存する度合いよりも強くなると考えられる。この索引語 w がある特定の記事（またはクラスタ） j に依存する度合いは、式(2.4.1)の $(\chi^2)^i_{wj}$ の分散値 $(\mathrm{var}(\chi^2)^i_w)$ で計算した。$(\mathrm{var}(\chi^2)^i_w)$ はその

値が大きいほど索引語 w が特定のクラスタ、または記事に強く依存することを示す。索引語 w のクラスタ（C）と記事（T）における文脈依存の度合いの関係を式(2.4.2)に示す。

$$(\mathrm{var}(\chi^2)_w^T) < (\mathrm{var}(\chi^2)_w^C) \tag{2.4.2}$$

式(2.4.2)において記事における索引語 w の分散値 $(\mathrm{var}(\chi^2)_w^T)$ よりもクラスタにおける索引語 w の分散値 $(\mathrm{var}(\chi^2)_w^C)$ が大きいことから、索引語 w は特定の記事よりも特のクラスタに強く依存することを示す。よって式(2.4.2)を満たす索引語 w を重要索引語と判定する。

2.5　本システムとの関連

本節では、前節までに説明した関連研究を踏まえて、本システムの特徴、違いを述べる。Scatter/Gather システムと Grouper システムの目的は、膨大な情報検索結果を意味的に揃ったグループにクラスタリングする事により検索結果の理解を容易にし、ユーザが求めている正解へ早く到達出来るような手助けをすることである。そのためには、単純なキーワード検索で得られる膨大な検索結果を絞り込む必要があり、Scatter/Gather システムでは、ユーザが選択したクラスタ内だけで再クラスタリングをすることにより、検索の範囲を絞り込む方法をとっている。Grouper システムでは、ユーザが選択したクラスタを特徴づけている索引語を、論理積をとる新たなキーワードの候補として提示することにより結果の絞込みをはかっている。このように、どちらのシステムも情報検索ツールとしての意味合いが強く、結果の可視化については、両システムとも通常の検索システムと同様の方法をとっており、表形式で表されたクラスタに含まれる特徴的な文書のタイトルや、索引語をリスト表示しているだけである。しかし、情報検索ツールに不可欠な検索速度の向上のため、Scatter/Gather システムでは、Buckshot 法と Fractionation 法を、Grouper システムでは、STC アルゴリズムを使用し、クラスタリングの精度を保ったまま速度の向上に努めている。このように Scatter／Gather システムと Grouper システムは答えが 1 つ存在するような検索モデルを想定した情報検索システムといえる。

これに対し、より複雑な問題－例えば環境問題など社会的な懸案事項－を個人が考えていく場合には、その問題にどのような経緯があったかなどの情報を収集し、自分なりの視点を作り上げていく必要がある。本システムは、このような問題に対処できるようなシステムとして設計されている。設計の指針としては以下に示す 2 つのポイントがある。

1. ユーザの視点を反映する機能として、視点を反映した情報を 1 つのグループに集める機能を盛り込む。
2. 結果の表示に時間軸を加え、情報の推移を可視化出来るようにする。

この 2 つのポイントを実現するシステムを構築することにより、単なる情報検索システムを越えた、個人の思考の整理を支援するシステムとなりうるのではないかと考えている。

第３章
社会情報可視化システムの構築

本研究の目的を達成する１つの例として、あるキーワードが属する文脈の可視化と、その文脈に個人的選好を反映できる機能を備えた実験システムを構築した。最初にシステムの概要と特徴を紹介し、次にシステムの動作アルゴリズム、操作方法を説明する。

3.1　システムの概要

システムの概念図を図 3.1 に示す。本システムは、ユーザーが関心のあるキーワードを選択すると、そのキーワードを含む新聞記事を、そのキーワードが属する文脈ごとにクラスタ分けして表形式で表示する。さらに、各々の記事の前文と全文を表示させたり、各クラスタの構造を時系列に沿って空間表示させることも出来る。システムによって自動生成されたクラスタが、ユーザーの分類と異なっている場合は、ユーザーの選好をシステムに与えることにより、その選好を反映するように再クラスタリングさせることも可能である。

索引語の抽出と、索引語ベクトルの算出、類似度の計算、初期クラスタリングは、実行速度を上げるため予め実施しておく。

今回、元文書群としては、CD-ROM 版の毎日新聞（1991 年から 1997 年）[毎日新聞]を用いた。文字処理は Perl、空間表示は Java 言語のアプレット、CGI は C 言語で実装した。

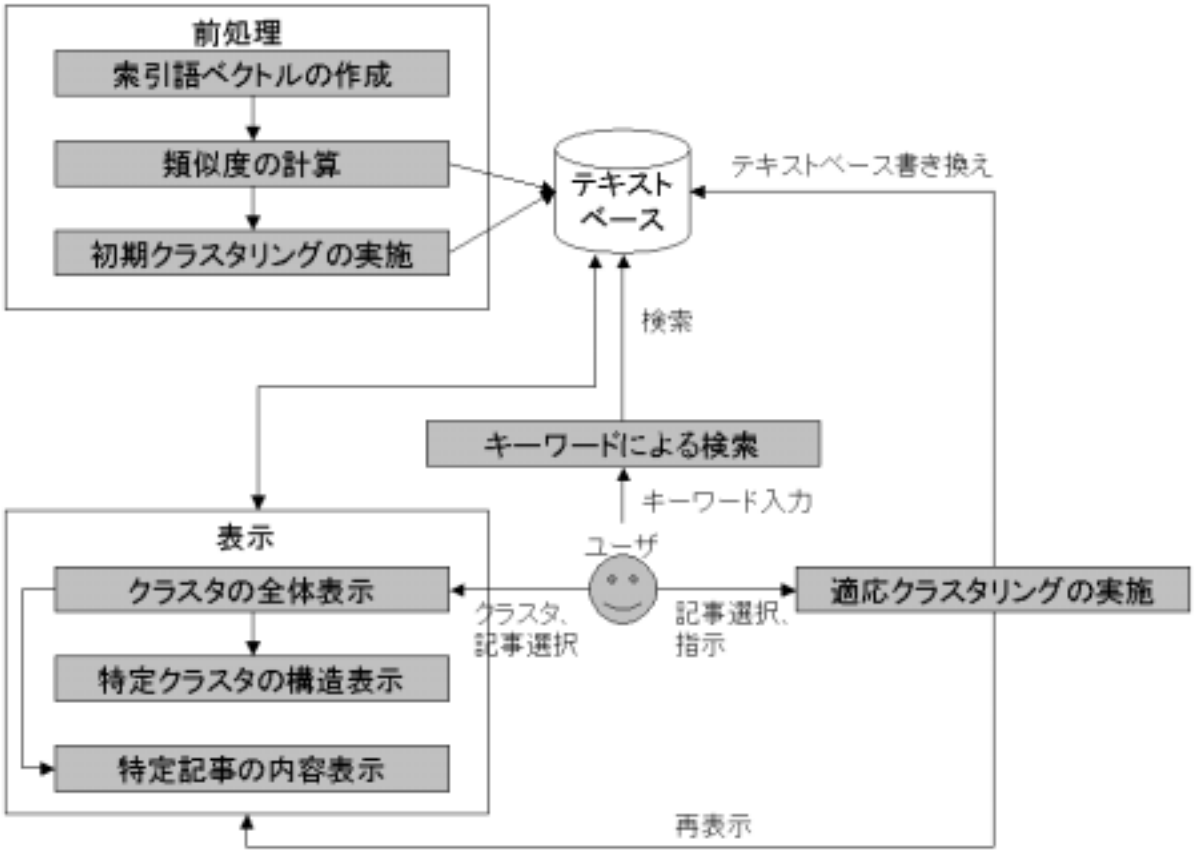

前処理
索引語ベクトルの作成
類似度の計算
初期クラスタリングの実施
テキスト
ベース
テキストベース書き換え
検索
キーワードによる検索
キーワード入力
表示
ユーザ
クラスタの全体表示
特定クラスタの構造表示
特定記事の内容表示
クラスタ、
記事選択
記事選択、
指示
適応クラスタリングの実施
再表示

図 3.1　システムの概要

3.2　システムの特徴

3.2.1　3 つのレベルでの文脈の可視化

本システムは、選択されたキーワードが属する文脈をクラスタとして表示することができる。文脈の表示には、文脈全体を表形式で表示する機能、特定文脈の構造を時系列上にグラフ表示する機能、特定の記事の内容を表示する機能の 3 つのレベルの表示形式が可能であり、社会情報の可視化を容易にしている。

3.2.2　適応クラスタリングによる個人的選好の反映

本システムが自動抽出した文脈は、多くの場合ユーザが考える文脈とは異なっている。これを補正するため、適応クラスタリングを実施し、ユーザの選好をシステムに反映することができる。この適応クラスタリングには、出来るだけユーザが選択した記事を同一クラスタに分類する再現率優先モードと、出来るだけごみを拾わない精度優先モードの 2 つのモードがあり、ユーザが選択できる。

この選好の反映は、同一セッション内ではシステム内に保持されるため、選好を追加的に反映させることも可能である。また、現在は同一セッション内でのみ個人的選好情報が保持できるが、この情報をファイルなどに記録することで、ユーザプロファイルとして管理可能である。

3.3　クラスタリング手法

本節では、クラスタリング手法を説明する。本システムでは、文書をクラスタリングするために以下のステップを踏む。

１．索引語の抽出
　　各文書から文書を特徴づける索引語を抽出する。

２．索引語ベクトルの算出
　　各索引語の索引語ベクトルを算出する。

３．類似度の計算
　　索引語ベクトルを用いて各文書間の類似度を計算する。

４．初期クラスタリングの実施
　　類似度を使ってクラスタリングを実施する。

５．適応クラスタリングの実施
　　ユーザーからのフィードバックを元に類似度の重みを変化させ、再クラスタリングする。

以下に各ステップで用いた手法を述べる。

3.3.1　索引語の抽出

索引語抽出の目的は、文書中からその文書を特徴づける索引語を漏れなく抽出することである。さらに抽出した索引語がその文書の内容にどれだけ密接に関連しているかを、索引語の重みとして付与する。重みとして次節で説明する索引語頻度－逆文書頻度値を用いる。

今回は元文書群として形態素解析済みの新聞記事を利用したが、その他ネットニュースの記事、Web 上のホームページなど様々なものが使用可能である。その場合は、前処理として「茶筌」などの形態素解析ツール[松本 99]を使って形態素解析し不要語を除く必要がある。

今回、本システムで新聞記事（毎日新聞）を利用したのは、形態素解析して不要語が取り除かれたデータが手軽に手に入れられること、情報の質が比較的均一なこと、幅広い社会情報を網羅していることを考慮してである。

3.3.2　索引語ベクトルの算出

索引語ベクトルの算出には、索引語頻度・逆文書頻度（TFIDF(Term Frequency Inverse Document Frequency)）を用いる。TFIDF とは、索引語の出現頻度(TF)に文書集合全体での索引語の出現の偏り(IDF)を考慮に入れて索引語の重要性を計算する方法である。単語の出現頻度に基づく重み付けの背景には、「何度も繰り返し言及される概念は重要な概念である」という仮説がある[Luhn 57]。この仮説に基づき、今回のシステムでは一つの文書に 2 回以上出現する索引語のみを対象とした。しかし、出現頻度は文書群全体の中での索引語の重要性を判断するのには利用できるが、ある特定の文書において索引語が重要であるかどうかの判断には利用できない。特定の文書における索引語の重要性を判断するためには索引語頻度の文書間の比較が必要となる。このような特定文書における索引語の重要性を表すための尺度として IDF が知られている。このように TFIDF は、文書中からその文書を特徴づける索引語を、比較的簡単に抽出する事が可能であり、また簡単でありながらも他の方法と同等にまたはそれ以上の性能を持つため本システムでも採用した。以下にその計算方法を説明する。

ある文書 D_r に t 個の索引語が出現するとする。１つの索引語 T_{ri} に単位ベクトル V_{ri} を対応させる。ベクトルを線形独立と仮定すると、t 次元のベクトル空間が定義される。t 個の索引語ベクトルの線形結合が、文書 D_r の内容を近似的に表現している。

式(3.3.2.1)

$$D_r = \sum_{i=1}^{t} a_{ri} V_{ri} \qquad (3.3.2.1)$$

ここで、a_{ri} は文書 D_r における索引語 T_{ri} の重みである。この重みを TFIDF で計算する方法を以下に説明する。

各文書で出現した索引語に通し番号をふる。ある文書 D_r の i 番目の索引語 T_{ri} の出現頻度 TF を tf_{ri} とする。次に、索引語 T_{ri} を含む文書数を文書頻度 df_{ri} とする。全文書数 N と文書頻度 df_{ri} の比の対数をとったものを idf_{ri} とする。

$$idf_{ri} = \log\left(\frac{N}{df_{ri}}\right) \tag{3.3.2.2}$$

ここで、出現頻度 tf_{ri} と idf_{ri} の積を索引語 T_i の重み w_{ri} とする。

$$w_{ri} = tf_{ri} \cdot idf_{ri} \tag{3.3.2.3}$$

となる。さらにこれを正規化する。

$$a_{ri} = \frac{w_{ri}}{\sqrt{(w_{r1})^2 + \ldots + (w_{rn})^2}} \tag{3.3.2.4}$$

この a_{ri} を文書 D_r における索引語 T_{ri} の重みとする。

3.3.3 類似度の計算

索引語の集合として表現されている文書間の類似度を表す尺度には、内積、Dice 係数（Dice coefficient）、Jaccard 係数(Jaccard coefficient)、余弦などの方法があるが、今回は単純でよく使われている各々の文書の索引語ベクトルの内積をとる方法を採用する。

この方法では、文書 D_r と文書 $D_s (= \sum_{j=1}^{t} q_{sj} V_{sj})$ の類似度 $sim(D_r, D_s)$ は式(3.3.3)のように索引語ベクトルの内積で表される。

$$\begin{aligned} sim(D_r, D_s) &= D_r \cdot D_s \\ &= \sum_{i=1}^{t} \sum_{j=1}^{t} a_{ri} q_{sj} \end{aligned} \tag{3.3.3}$$

3.3.4 クラスタリング手法

クラスタリングを実施するには類似度の計算法と、クラスタリング手法を決める必要がある。文書間の類似度の計算法は前節で説明した。クラスタ間の類似度の計

算方法には、単連結法(single-link method)、完全連結法(complete-link method)、群間平均法(group average method)、重心法(centroid method)、ウォード法(Ward method)などがある。このうち単連結法では、ある文書と最も類似度の高い文書をとの類似度をクラスタ間の類似度とし、逆に完全連結法では、最も類似度の低い文書との類似度を採用する。よって、一組の文書間の類似度だけが大きな影響を与えることになる。この問題を補う方法が、本システムで採用した群間平均法である。郡間平均法では、クラスタ C_h と C_l の類似度は、式(3.3.4)により計算される。

$$sim(C_h, C_l) = \frac{\sum_{D_j \in C_h} \sum_{D_k \in C_l} sim(D_j, D_k)}{N_h N_l} \qquad (3.3.4)$$

クラスタリングの方法には、階層型クラスタリング(hierarchical clustering)と非階層型クラスタリング(non-hierarchical clustering)がある。階層型クラスタリングは、クラスタリング結果をデンドログラムとして得られるため、詳細な解析が必要な場合に向いているが、計算量は非階層型より大きくなる。今回は、データセットの大きさを考慮し、計算量が小さい非階層型クラスタリングを採用する。非階層型クラスタリングの方法には、いくつかの種(seed)を決めて、その種のどれかに全ての文書を所属させる K-means 法が用いられることが多い[Douglass 92]が、いくつ種を設定すればよいのかが明確ではない、どの文書を種にすればよいか分からない問題がある。本論文では最も単純な単一パス法を採用する。その手順は、

① 最初の文書を、最初のクラスタとする。
② 次の文書と、その時点で存在する全てのクラスタの代表を群間平均法で照合する。
③ 一致の程度にしたがって、その文書をいずれかのクラスタに割り当てる。もし一致の程度が閾値（本システムでは 0.0001）に達しなければ、それを代表とする新しいクラスタを生成する。
④ 手順②に戻る。

3.3.5　適応クラスタリング手法

適応クラスタリング方法については第４章、第５章で述べる。

本システムでは、適応クラスタリング手法として精度優先と再現率優先のいずれかのモードを選択できる。精度優先モードの場合は、第５章で説明するように、TFIDF 値を使った方法で重要索引語の割合を 56.7%、重みを 5 倍に設定してある。再現率優先モードの場合は、見なし共起を使った方法で、重要索引語の割合を 41.7%、重みを無しに設定してある。

3.4　ユーザインタフェイス

本節では、ユーザインタフェイスを説明する。本システムでは、クラスタリング結果を３種類のレベルで表示させることが可能である。

１．　全体結果表示

ユーザーが選択したキーワードを含む記事のクラスタリング結果を表形式で表示する。

２．　特定記事の内容表示

ある特定の記事の前文と全文を表示する。

３．　特定クラスタの構造表示

ある特定のクラスタに属する記事の構造を時系列上に表示する。

以下に全体を通した操作方法と結果の表示方法の詳細を述べる。

3.4.1　キーワード選択画面

本システムを起動すると、最初に図 3.4.1 に示すようなキーワード選択画面になる。ユーザーは表示されているキーワードの内、どれか１つを選択する。本来であれば、通常の検索エンジンのようにユーザにキーワードを自由に入力させることが望ましいが、今回は計算時間の節約のため、あらかじめ前処理をしておいたキーワードのみ選択できるようになっている。

図 3.4.1　キーワード選択画面

3.4.2　全体結果表示画面

キーワード選択画面でキーワードを選択すると、全体結果表示画面になる。この画面では図 3.4.2 に示すように、各々のクラスタに属する記事が表形式で表示される。これはユーザがキーワード選択画面で選択したキーワードを含む記事の分類（クラスタ）と、その分類に含まれる記事の一覧である。表の左側のカラムに記事のタイトルが、右側のカラムにその記事を構成する索引語が表示される。本システムのこれ以降の操作は全てこの全体結果表示画面から実行する。

図 3.4.2　全体結果表示画面

3.4.3　特定記事の内容表示

全体結果表示画面上の記事のタイトルは、その記事の前文と全文を表示する画面へのハイパーリンクとなっている。ある記事のタイトルを選択した結果を図 3..4.3 に示す。

図 3.4.3　特定記事の内容表示画面

3.4.4　特定クラスタの構造表示

全体結果表示画面により記事の分類の一覧を、特定記事の内容表示画面により記事の要約（＝前文）と全文を表示することが可能である。しかし、これだけでは各々のクラスタが表している主題と構造を読み取るのは困難である。クラスタの主題を表現するにはクラスタの内容を何らかの方法で要約する必要がある。クラスタの要約方法としては、クラスタ中で出現する頻度の高い単語やフレーズに加え、典型的なタイトルを表示する方法がとられている[Zamir 99] [Hearst 95]。しかし、この方法ではクラスタを構成している記事の生成日時を考慮していないため、情報の発生時間を含む全体構造が表現できない。本システムでは、全体結果表示画面の索引語の表示によりクラスタの主題を表現させることとし、さらに特定のクラスタの全体構造を表現する方法として、クラスタを構成する記事とそれを特徴づける索引語のグラフ表示を行う。

このグラフ表示の目的は、記事の生成月日の分布と、記事を構成している索引語の可視化である。

グラフ自動描画の基本的枠組みとして、描画対象、描画規約、描画規則、優先関係、描画アルゴリズムの５項目を考えることが必要である[杉山 93]。

描画対象となるグラフは、木、有向グラフ、無向グラフ、複合グラフにクラス分類される。描画規約は頂点と辺の配線に間する基本約束であり、描画に際し必

ず満たされるべき制約である。描画規約は、頂点の配置規約および辺の配線規約からなる。頂点の配置規約には自由配置、平行線配置などがあり、辺の配線規約には直線配線などの線種と座標系との関係の 2 種類がある。描画規則は出来る限り満たすべき制約であり、ひとつの描画において満たすことが望ましい規則である静的規則と、グラフを変えていくときに、連続する複数の描画の間に成立する規則である動的規則の 2 つがある。静的規則は意味的規則と構造的規則に分けられる。意味的規則は、頂点や辺の意味からくる配置・配線規則であり、構造的規則はグラフの構造情報だけに関係した規則である。描画規約と描画規則をまとめて美的基準とも呼ぶ。

本システムでの描画対象は、クラスタを構成している記事とその記事を構成する索引語であり、今回はその特性を考慮し平面無向グラフとし、以下にその美的基準をまとめる。描画規則は優先順位の高いものから示す。

- 描画規約
  - † 記事を表す頂点は赤い長方形で描き、平行線上に配置する。平行線は時間軸と直交しており、記事の生成時間を表す平行線上に配置する。
  - † 記事を構成する索引語は黄色の長方形で描き、自由配置である。
  - † 記事とその記事を構成する索引語は直線で配線する。
  - † 配線は座標系の線に独立である。
- 描画規則
  - † 記事とその記事を構成する索引語間の配線の長さは、その索引語の重みに反比例させる。
  - † 頂点どうしの重なりは出来るだけ減らす。
  - † 配線どうしの重なりは出来るだけ減らす。

最後に本システムで採用した描画アルゴリズムを説明する。平面無向グラフにおける描画アルゴリズムについては、力学モデルに基づいたマグネティック・スプリング・モデル[三末 94]など様々な手法が考案されている。本システムでも、これらの手法を参考にした方法をとるが、本グラフ表示の目的と美的基準を鑑みて、描画の精度よりも描画速度を優先した単純なアルゴリズムを採用する事とし、頂点間のスプリング力やエネルギーは考慮しない。

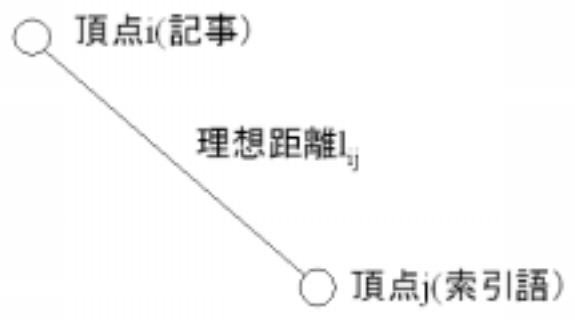


図 3.4.4.1　頂点間の理想距離

まず、図 3.4.4.1 に示すように、記事を表す頂点 i とその記事に属する索引語を表す頂点 j の間の理想距離 l_{ij} は、その索引語の類似度 R_{ij} を使って式(3.4.4)のように算出しておく。

$$l_{ij} = m\frac{1}{R_{ij}} \tag{3.4.4}$$

ここで m は任意の定数で、空間配置しようとする画面の大きさで調節する。

次に、全ての頂点は、その記事の生成日時を表す平行線上にランダムに配置され、関連のある頂点間の距離を理想距離に近づける最適化と、関連の無い頂点間の距離を長くする最適化が行われる。また、同時に全頂点を画面の中心に近づける最適化も行われる。この 3 つの最適化のステップを繰り返すことにより、本システムの美的基準を満たす準最適解としての描画が完成する。

本システムにおいて、ある 1 つのクラスタの構造を表示するには、全体結果表示画面上のクラスタ番号の横にある構造表示を選択する。図 3.4.4.2 に示すように、そのクラスタに属する記事が、記事の発行日時をｘ軸として表示される。さらにその記事を構成する索引語が記事のまわりに配置される。

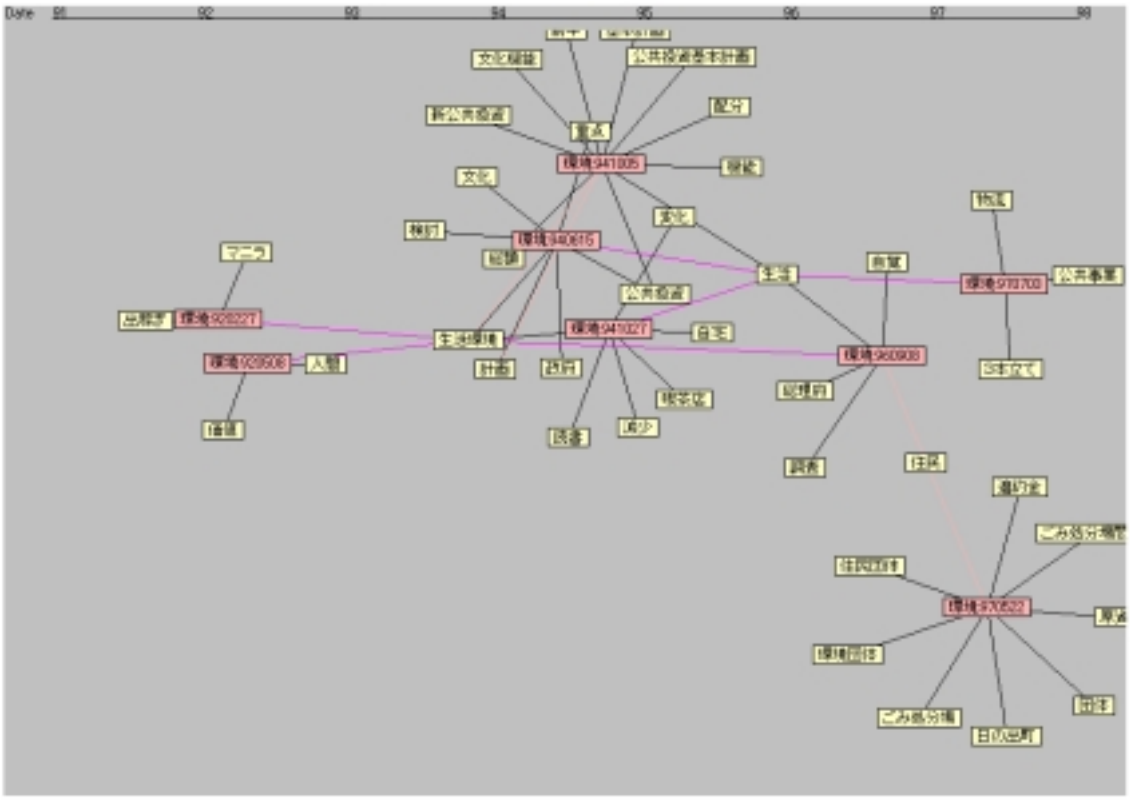


図 3.4.4.2　特定クラスタの構造表示画面

3.4.5　適応クラスタリング手法

適応クラスタリングを実施するには、図 3.4.5.1 に示すように、全体結果表示画面の各記事の先頭にあるチェックボックスを選択し、Submit ボタンを押す。この際、再現率と精度のどちらを優先するかを選択することができる。結果は図 3.4.5.2 に示すようになる。ユーザーが選択した記事は赤字の selected が付与される。

図 3.4.5.1　適応クラスタリングの選択方法

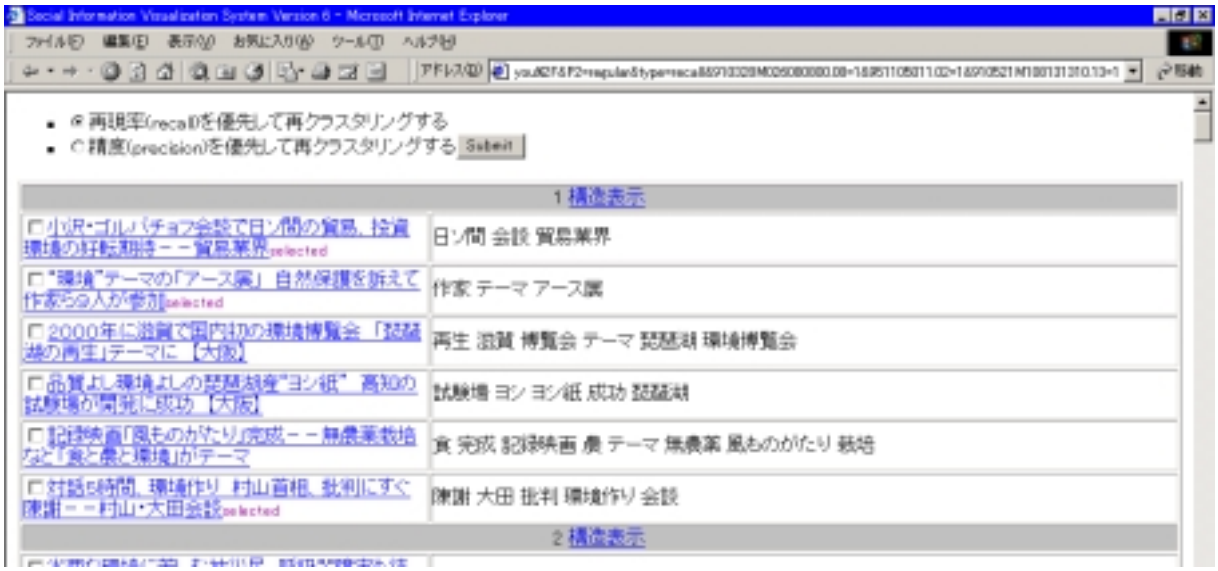


図 3.4.5.2　適応クラスタリング結果画面

3.5　実行結果

本節では、本システムの実行結果を例を用いて説明する。

図 3.5.1 は「環境」を選択した場合の全体結果表示画面である。この例において、クラスタ 51 中には索引語として‘ CO2 ’[1]、‘ 温室効果ガス ’などが多く出現しており、さらに「CO2 排出を抑制　環境面の目標を加える――電力需要見通し」、「日本の温室ガス削減 5%案、国会議員の 8 割「不十分」――環境 NGO アンケート」などの記事の見出しをみても、このクラスタは CO2 による地球温暖化を表していることが読み取れる。しかし、他のクラスタ内にも地球温暖化に関する記事があるので、それを全て選択し、再現率優先の適応クラスタリングを実施した。その結果を図 3..5.2 に示す。この例では、選択した記事のほとんどが１つのクラスタにまとまっている。またユーザは選択しなかったものの、それに関連の深い記事も同じくラスタに分類されている。

さらに、このクラスタが表している内容の発生時間の分布を見るためには、構造表示画面を選択して表示させればよい。図 3.5.3 はこの例におけるクラスタ 51 の初期構造表示である。図 3.5.4 は適応クラスタリング後の構造表示である。この例では再現率を優先させたので、適応クラスタリング後の記事の個数が比較的多くなり、一つ一つの記事に関連する索引語を判別することは難しい。しかし、記事全体の時間軸上の分布は十分に見て取ることが可能であり、この例では、地球温暖化に関する関心は 1996 年頃から急速に高まってきていることが窺える。

---

[1] 電子化した毎日新聞はすべての文字が全角文字に直されているので、CO_2 は CO2 と表記されている。

図 3.5.1　初期全体結果表示画面（環境－地球温暖化）

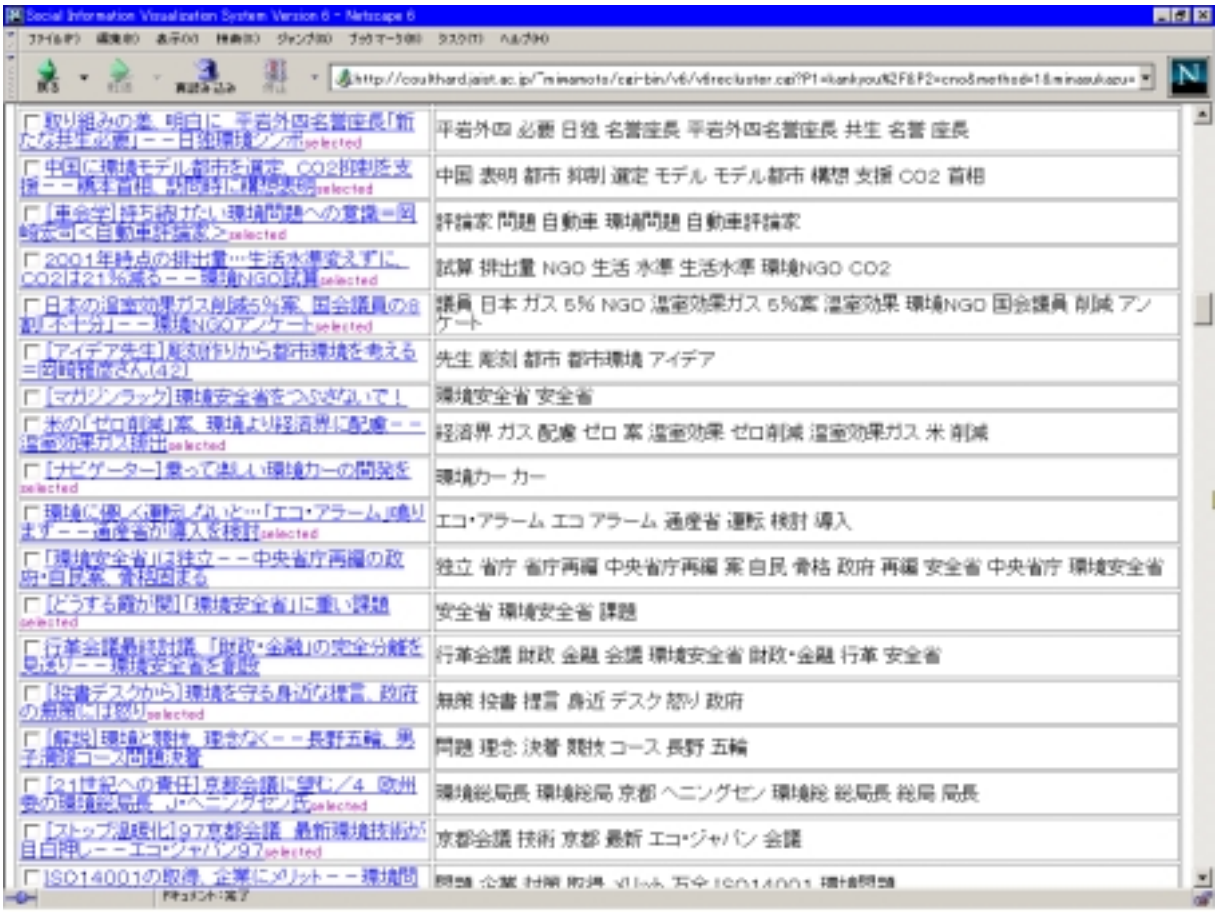


図 3.5.2　適応クラスタリング後の全体結果画面（環境－地球温暖化）

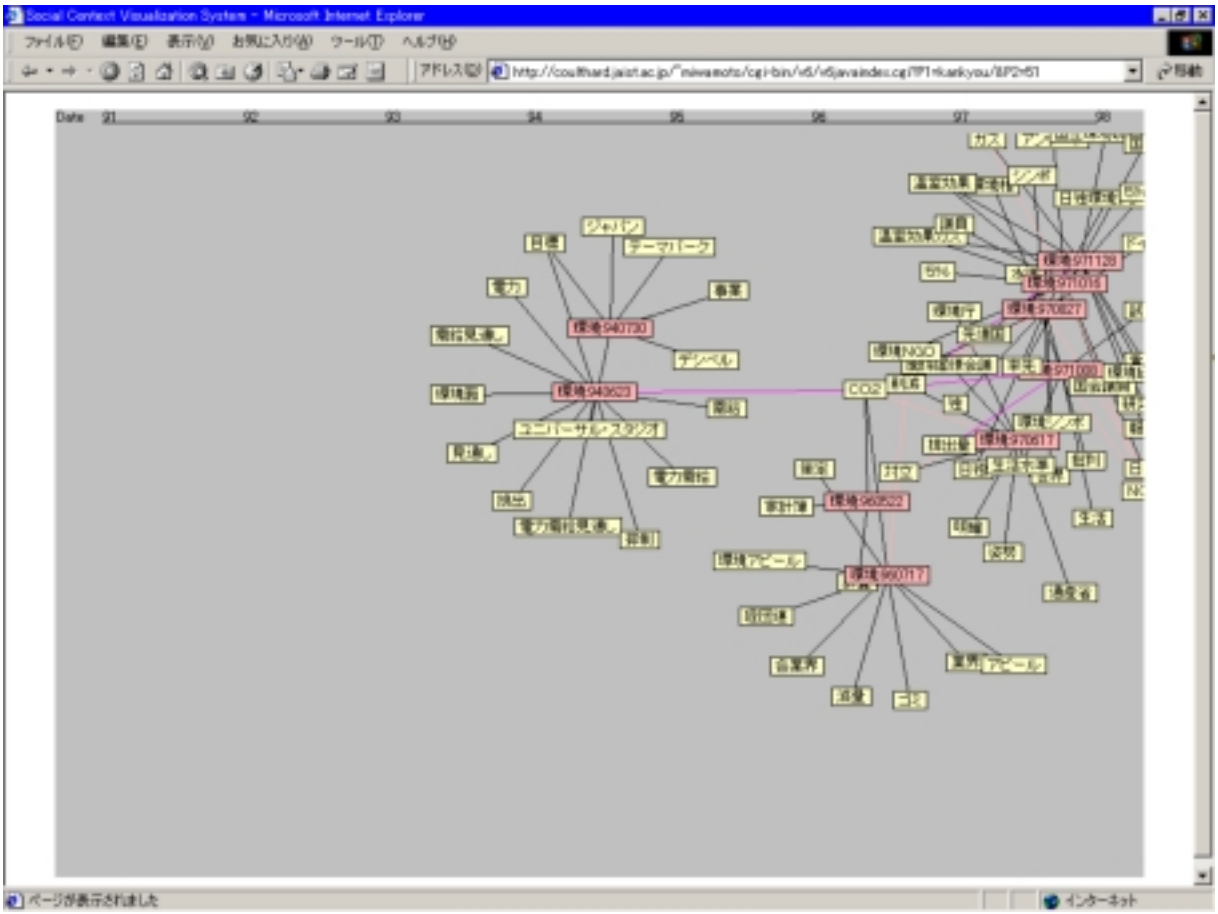


図 3.5.3　初期構造表示画面（環境－地球温暖化）

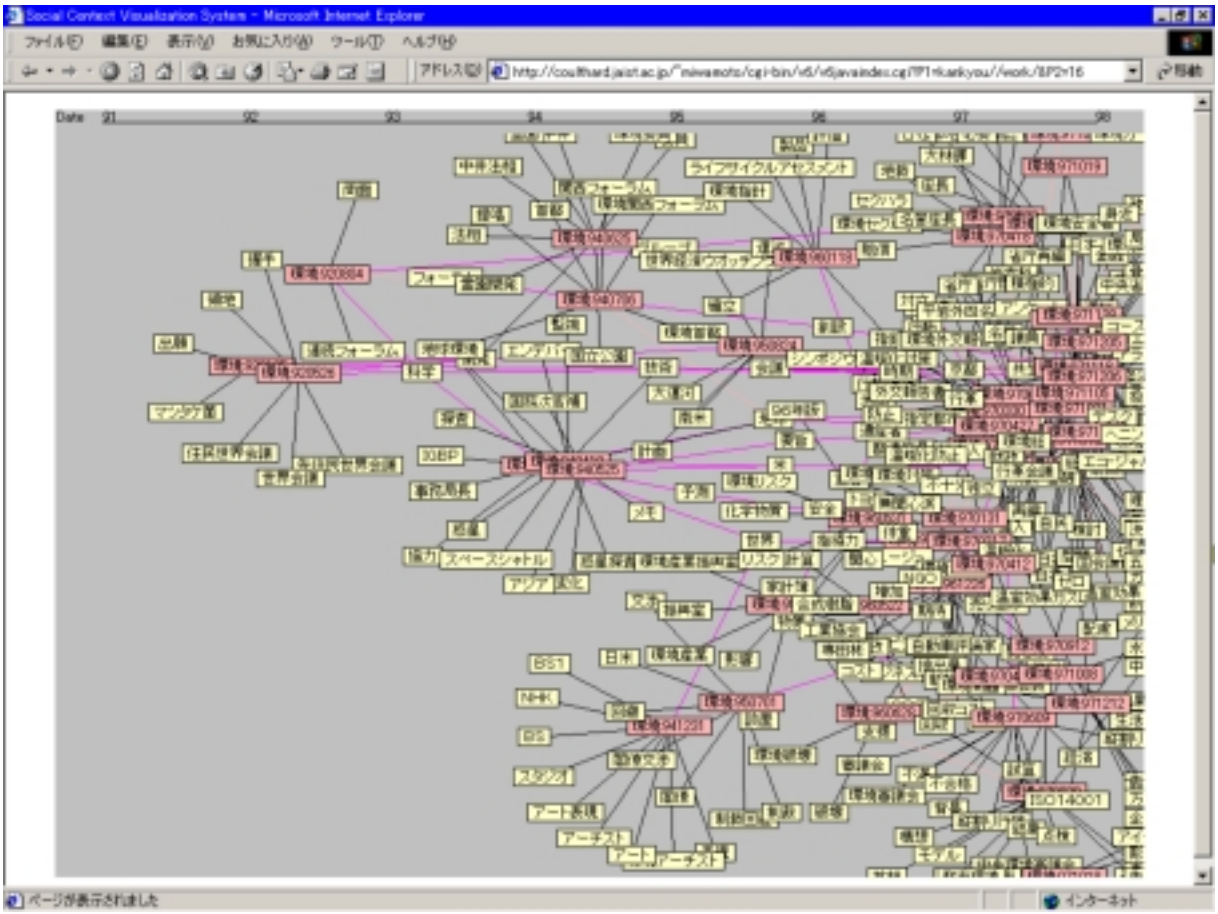


図 3.5.4　適応クラスタリング後の構造表示画面（環境－地球温暖化）

第４章
適応クラスタリング

本章では、最初に本研究での適応クラスタリングの概念について説明し、その具体的方法である索引語－逆文書頻度値、$\_^2$ 値、文脈依存の度合い、見なし共起を使った適応クラスタリングについて述べる。

4.1　適応クラスタリングの概念

第 2 章の関連研究でも説明したように、情報検索システムの膨大な検索結果の可視化方法の 1 つとしてクラスタリングを使った方法が広く研究されている。しかし、クラスタリングはどのようなアルゴリズムであっても、対象となる文書を索引語によって代表させる近似的な方法であり、その索引語間の類似度でどのクラスタに属させるかを決めている。これを‘正しい’または‘適切な’クラスタにするにはどうすればよいか考えてみよう。ここで‘正しい’クラスタ、‘適切な’クラスタという表現を使ったが、この正しさ、適切さを判断できるのはそのクラスタを使うユーザだけであり、そのユーザの持つ興味、視点によって判断は変わってくる。万人に共通する正しさ、適切さは存在しない。

そこで、ユーザの持つ興味、視点をクラスタリングに反映させることが必要になる。その方法として、図 4.1 のように、システムが提供する初期クラスタにユーザの視点を反映するためにフィードバックを行い、そのフィードバック情報に基づいたクラスタリングをする必要がある。このユーザからのフィードバック情報に基づいたクラスタリングのことを適応クラスタリングと呼ぶ。

この適応クラスタリングの実施には、ユーザからどのようなフィードバック情報を得るかと、その結果システムの何を変更するかの 2 点が必要となる。本研究では、

クラスタリング結果の適合性の判定を、ユーザから得るフィードバック情報とする。具体的には、ユーザはシステムが提示するクラスタを見て、自分自身の視点で同じ意味に属していると判断する記事を選択しシステムにフィードバックする。

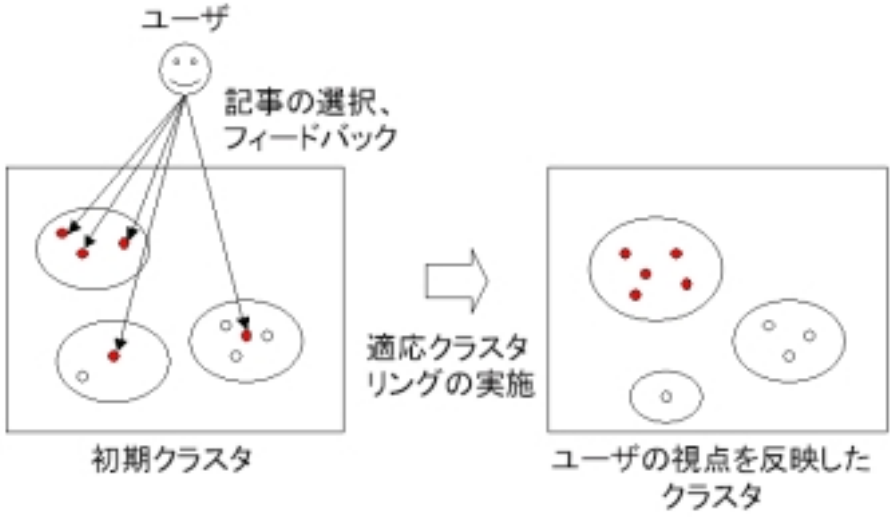


図 4.1　適応クラスタリングの概念図

次に、クラスタを変化させるには索引語の重みを変化させる。この場合、どの索引語を重要と見なすかが重要となる。重要索引語の決定方法については、4.2 節から 4.4 節で説明する。

ここで、ユーザからのフィードバック情報を使って文書の索引語の重みを修正する方法を式(4.1)で説明する。

$$T_{i+1} = \alpha T_i \tag{4.1}$$

$$\text{ただし、} \begin{array}{ll} 1 < \alpha & if \quad D_i^+ \\ 0 < \alpha < 1 & if \quad D_i^- \end{array}$$

式(4.1)の T_i は i 回目の検索質問に対応する文書の索引語ベクトルであり、T_{i+1} はそれを修正した文書の索引語ベクトルである。D_i^+, D_i^- はぞれぞれ、個々の適合文書、不適合文書に対応する索引語のうち、重み付けをする索引語（重要索引語と呼ぶ）である。α は索引語の重みに掛ける係数であり、適合文書、不適

合文書をどれくらい重要視するかを調節する。この方法では、重要索引語と重みを決定する必要がある。

文書中に共起する索引語がない場合でも、お互いの文書の意味は似かよっていて同一クラスタに分類したい場合もある。このような場合、文書の索引語の共起する度合いによるクラスタ分けで同一クラスタに分類するのは原理的に不可能であり、文書を構成する索引語の重みを増減しても意味は無い。

そこで、同一クラスタに分類したい文書に存在する索引語のうち、別の文書に属する索引語どうしを共起したと見なす（見なし共起と呼ぶ）方法を提案する(4.5 節)。

4.2　TFIDF 値を使った適応クラスタリング

ユーザからのフィードバックとして適合文書が得られた場合に、重要索引語を TFIDF 値を使って求める方法を説明する。この方法は、式(4.1)の重要索引語D_i^+を決めることに相当する。

まず、ユーザからフィードバックされた適合文書に含まれる索引語を抽出し、各々の TFIDF 値を求める。TFIDF 値の計算には様々な方法がある[Salton 88]が、本研究で採用した方法を以下に説明する。

各テキストで出現した索引語に通し番号をふる。あるテキストD_rの i 番目の索引語T_{ri}の出現頻度 TF をtf_{ri}とする。次に、索引語T_{ri}を含むテキスト数を文書頻度df_{ri}とする。全文書数Nと文書頻度df_{ri}の比の対数をとったものをidf_{ri}とする。

$$idf_{ri} = \log\left(\frac{N}{df_{ri}}\right) \tag{4.2.1}$$

ここで、出現頻度tf_{ri}とidf_{ri}の積を索引語T_iの重みw_{ri}とする。

$$w_{ri} = tf_{ri} \cdot idf_{ri} \tag{4.2.2}$$

となる。さらにこれを正規化する。

$$a_{ri} = \frac{w_{ri}}{\sqrt{(w_{r1})^2 + \ldots + (w_{rn})^2}} \tag{4.2.3}$$

このa_{ri}をテキストD_rにおける索引語T_{ri}の TFIDF 値とする。

この TFIDF 値は適合文書に含まれる索引語から重要索引語を決めるためにだけ使われる。つまり、適合文書に含まれる索引語を TFIDF 値で降順で並び替え、上位から適当な割合の索引語を重要索引語とする。この重要索引語に、重みαを加え、クラスタリングする。これが TFIDF 値を使う適応クラスタリングである。重要索引語となる索引語の最適な割合と、重みαは実験的に決定する。

4.3　χ^2値を使った適応クラスタリング

ユーザからのフィードバックとして適合文書が得られた場合に、重要索引語をχ^2値を使って決定する方法を説明する。この方法は、式(4.1)の重要索引語D_i^+を決めることに相当する。

最初に、χ^2値を使った重要語の抽出について説明する。一般にχ^2検定とは、2 つの要因の独立性を検定する方法である。ここでは、要因として記事内における索引語の出現頻度と、クラスタ内における索引語の出現頻度を採用する。もし、ある索引語が全記事中で全く均質に分布している場合は、2 つの要因は全く無連関の状態にあり、χ^2値は最小値０となる。逆に、ある検索語がある特定のクラスタ内にしか出現しない場合は、最大連関の状態にありχ^2値は最大値をとる。よって、χ^2値により、ある索引語の記事内とクラスタ内の出現頻度の偏りを計算することが出来る。このχ^2法を用いて重要なキーワードを抽出する研究が行われており[長尾 76]、χ^2法がキーワードの抽出に有効であることが確かめられている。また、重要漢字の自動抽出にχ^2法を用いる研究もされている[渡辺 94]。しかし、χ^2値をそのまま使う方法では、記事数の多いクラスタのχ^2値は大きく、逆に記事数の少ないクラスタのχ^2値は小さくなってしまう。そこで、[渡辺 94]が採用したそれぞれのクラスタにおける出現頻度の理論度数からのずれに着目する方法を、本研究でも採用した。

以下に本研究で用いた方法を説明する。検索語wがi（iは記事またはクラスタ）において特定の記事（またはクラスタ）jに依存する度合いを式(4.3)に示す。

$$(\chi^2)_{wj}^{i} = \begin{cases} \dfrac{(x_{wj} - m_{wj})^2}{m_{wj}} & if \quad x_{wj} > m_{wj} \\ 0 & otherwise \end{cases} \tag{4.3}$$

ここで、

$$m_{wj} = \frac{\sum_{j=1}^{n} x_{wj}}{\sum_{w=1}^{m}\sum_{j=1}^{n} x_{wj}} \times \sum_{w=1}^{m} x_{wj}$$

ただし、

i　　記事、またはクラスタ

m　　索引語の数

n　　記事、またはクラスタの数

x_{wj}　　特定の記事またはクラスタ j における索引語 w の出現頻度

m_{wj}　　特定の記事またはクラスタ j における索引語 w の理想頻度

理想頻度とは、全記事に等確率でその索引語が出現した場合の出現頻度である。

まず、ユーザからフィードバックされた適合文書集合を新たなクラスタとし、現在のクラスタからユーザがフィードバックした記事を除いたものを、それぞれ新たなクラスタと見なす。このクラスタと記事を対象として、ユーザからフィードバックされた記事中の索引語の χ^2 値を式(4.3)で計算する。この χ^2 値は適合文書に含まれる索引語から重要索引語を決めるためにだけ使われる。つまり、適合文書に含まれる索引語を χ^2 値で降順で並び替え、上位から適当な割合の索引語を重要索引語とする。この重要索引語に、重み α を加え、クラスタリングする。これが χ^2 値を使う適応クラスタリングである。

重要索引語となる索引語の最適な割合と、重み α は実験的に決定する。

4.4　文脈依存の度合いを使った適応クラスタリング

ここでは、ユーザからのフィードバックとして適合文書が得られた場合に、重要索引語を文脈依存の度合いを使って決定する方法を説明する。この方法は、重要語と一般語を抽出するので式(4.1)の重要索引語D_i^+, D_i^-を決めることに相当する。

文書の自動分類において、文脈依存の度合いを用いて重要語を抽出する研究が行われている[福本 99]。本研究でも、この研究を基にして文脈依存の度合いを算出した。

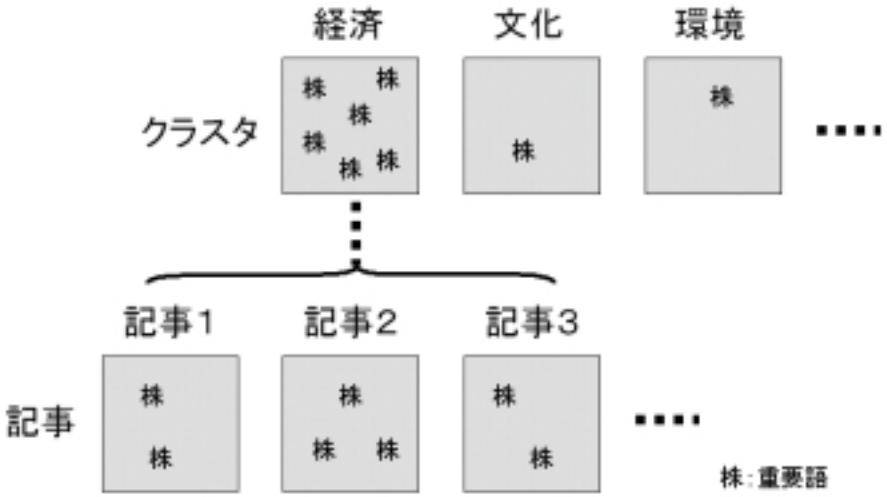


図 4.4　新聞記事の構造　（[福本 99]より一部変更の上、転載）

図 4.4 で示される新聞記事の構造において、文脈依存の度合いとは、ある索引語がこの図で示した特定のクラスタ、あるいは特定の記事とどのくらい深く関っているかという度合いの強さを示す。例えば、図 4.4 において、‘ 経済 ’ に関するクラスタにおける重要索引語を ‘ 株 ’ とすると、‘ 株 ’ は各記事にまたがり出現する。よって ‘ 株 ’ は各記事での分布の偏りが一般語と同様、小さく、特定の記事に依存する度合いは低い。次に ‘ 経済 ’ のクラスタについて考える。一般語は色々な文書中に均等に現れるため、各クラスタにおける分布の偏りと、記事における分布の偏りに差はない。一方、‘ 株’の ‘ 経済’での依存の度合いは ‘ 株 ’ が ‘ 経済 ’ という特定のクラスタに集中して出現するため、結果的に特定の記事に依存する度合いよりも強

くなると考えられる。

この索引語 w がある特定の記事（またはクラスタ） j に依存する度合いは、式(4.3)の $(\chi^2)^i_{wj}$ の分散値 $(\mathrm{var}(\chi^2)^i_w)$ で計算した。$(\mathrm{var}(\chi^2)^i_w)$ はその値が大きいほど索引語 w が特定のクラスタ、または記事に強く依存することを示す。索引語 w のクラスタ（C）と記事（T）における文脈依存の度合いの関係を式(4.4)に示す。

$$(\mathrm{var}(\chi^2)^T_w) < (\mathrm{var}(\chi^2)^C_w) \tag{4.4}$$

式(4.4)において記事における索引語 w の分散値 $(\mathrm{var}(\chi^2)^T_w)$ よりもクラスタにおける索引語 w の分散値 $(\mathrm{var}(\chi^2)^C_w)$ が大きいことから、索引語 w は特定の記事よりも特のクラスタに強く依存することを示す。よって式(4.4)を満たす索引語 w を T_i^+ 、それ以外の索引語を T_i^- とした。

重要語と一般語の重みは、トレーニングデータが存在する場合は、それを使った学習によって求めることも出来る[福本 99]。本研究では、予めトレーニングデータを用意することは出来ないので、重み α は実験的に決定する。

4.5　見なし共起を使った適応クラスタリング

図 4.5 の例で、文書 1 と文書 2 が同一クラスタに分類したい文書としてユーザからフィードバックされたとする。この場合、索引語 01 は両方の文書に含まれているので通常の共起語として扱われるが、各々の文書に独立して存在する索引語 02 と索引語 03、索引語 04 をそれぞれ共起したと見なし、クラスタリング時のテキスト間の類似度の計算に使用する。例えば、文書 3 と文書 4 の間の類似度を計算する場合、共起している索引語 10 だけでなく索引語 02 と索引語 03 についても共起したと見なし、文書間の類似度の計算をする。

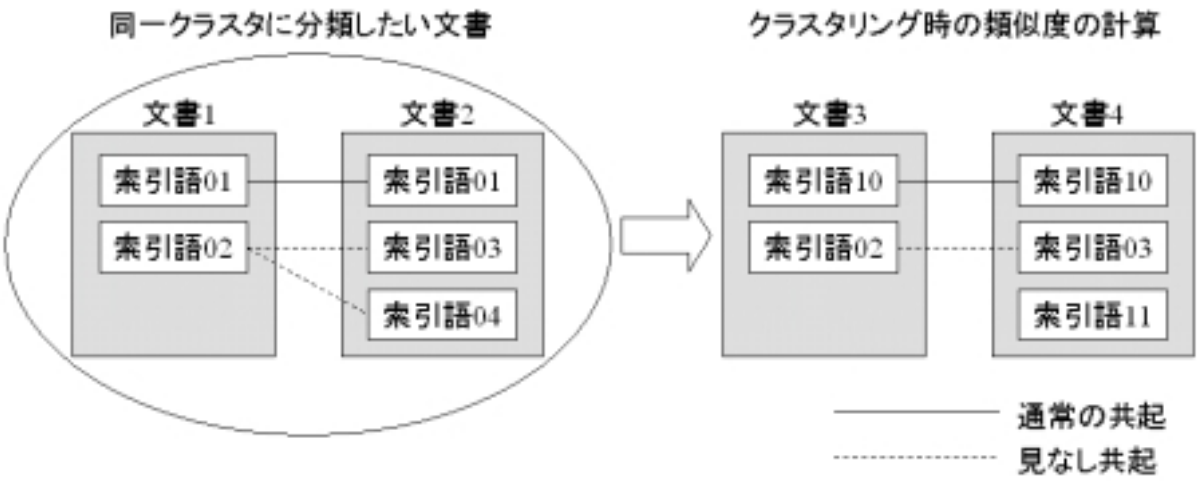


図 4.5　見なし共起を使った類似度の計算

ただし、該当する全ての索引語を共起したと見なすと、過適応になる恐れがあるので、該当する索引語の TFIDF 値の上位から適当な割合の索引語だけを重要索引語として使用することにする。索引語の最適な割合は実験的手法で決定する。

第５章
適応クラスタリング手法の評価

本章では、適応クラスタリング手法の評価について、その実験方法を記述し、結果を考察する。

5.1　データ

実験で用いたデータは、1991 年から 1997 年までの毎日新聞の内、文字数 400 字から 999 字の全 202,667 記事を使った。記事を長さの観点から見ると、非常に短い記事は死亡記事、人事記事など、非常に長い記事は選挙の結果、国立大学の試験要綱など、羅列に近い記事である傾向が見られるとされる[豊浦 97]。この調査をもとに、社会情報の対象として実質的な内容を持つ可能性の高い、文字数 400 字から 999 字の記事をデータとして用いた。

毎日新聞は分野分けがなされていないので、適当なキーワードを含む記事を抽出したのち、人手により分類しこれを正解セットとする。通常、人手による分類では 1 つのキーワードを含む記事は、その意味や使われ方により 10 個から 50 個程度に分類される。本実験で使用したキーワードと正解セット中の 1 つの分野を、表 5.1 に示す。

この表の、キーワード欄の記事数とは、キーワードが 2 回以上出現する記事の総数である。「正解セット中の共起索引語数／正解セット中の全索引語数」は、正解セット中の記事の索引語の中で共起している割合であり、正解セット中の記事がどのくらい同じ索引語で構成されているかを示している。また、「正解セット中の全索引語数／キーワードを 2 回以上含む全記事中の全索引語数」は、クラスタリングのベースとなる記事集合中の全索引語とユーザがフィードバックした記事中の索引語の

割合である。

キーワード (記事数)	正解セット (記事数)	*1	*2
核(293)	核抑止力(14)	60/202(30%)	202/5420(3.7%)
原子力(246)	原発反対(12)	46/132(35%)	132/2870(4.5%)
環境(330)	地球温暖化(27)	196/474(41%)	474/5209(9.1%)
福祉(229)	政策(40)	204/503(41%)	503/2671(18.8%)
通信(407)	無線(13)	54/106(51%)	106/4348(2.4%)
情報(614)	情報開示(42)	426/826(52%)	826/9492(8.7%)

*1:正解セット中の共起索引語数／正解セット中の全索引語数

*2:正解セット中の全索引語数／キーワードが 2 回以上現れる全記事中の全索引語数

表 5.1　キーワードと正解セットの特性

5.2　実験方法

最初に、実験の手順を説明する。

1. 1 つのキーワードを含む記事の集合を、第 3 章で説明した方法を使ってクラスタリングし、初期クラスタを求める。
2. 正解セットから 1 つの分野を選択し、その分野中の記事をユーザからのフィードバック情報として、適応クラスタリングを実施する。
3. その結果得られたクラスタのうち、選択した正解セットの分野に最も近いクラスタを選び、正解セットとの違いを再現率、精度とその要約である f 値で比較する。
4. 重要索引語の重みを 1.5、2、5、10 倍にとり、それぞれの場合において重要索引語の割合を TFIDF 値（χ^2 値）の上位から 0%から 100%まで 10%きざみで変化させて実験する。

本研究での再現率(recall)、精度(precision)、f 値(harmonic mean of recall and precision)の定義をそれぞれ式(5.2.1)、式(5.2.2)、式(5.2.3)に示す。

$$\text{再現率} = \frac{\text{あるクラスタ中の正解記事数}}{\text{正解セット中の記事数}} \tag{5.2.1}$$

$$\text{精度} = \frac{\text{あるクラスタ中の正解記事数}}{\text{あるクラスタ中の記事数}} \tag{5.2.2}$$

$$\text{f 値} = \frac{2}{\frac{1}{\text{再現率}} + \frac{1}{\text{精度}}} \tag{5.2.3}$$

5.3　結果と考察

本節では、前述の 4 種類の適応クラスタリング方法の性能について、評価実験の結果とその考察を記述する。

5.3.1　TFIDF 値を使った方法の結果と考察

実験結果を、図 5.3.1.1～図 5.3.1.6 に示す。‘ 核－核抑止力 ’ の場合は、重要索引語の割合が 70%～90%で f 値が小幅に改善する傾向が見られる。‘ 原子力－原発反対 ’、‘ 環境－地球温暖化 ’、‘ 通信－無線 ’ の場合は、f 値はほとんど変化しないか、むしろ悪化している。‘ 福祉－政策 ’ の場合は、10%～80%程度にとった時に f 値の小幅な改善が見られた。‘ 情報－情報開示 ’ の場合は、20%～50%程度にとった時に f 値が改善する。

いずれの場合も、再現率と精度の両方が f 値の改善に寄与している。再現率はいずれの場合も小幅な改善に留まっており、ユーザからフィードバックされた記事を 1 つのクラスタにまとめる効果はほとんど期待できない。精度については、f 値の改善が見られた ‘ 核－核抑止力 ’、‘ 福祉－政策 ’、‘ 情報－情報開示 ’ のいずれの場合も重みを 5 倍にとった時に最も精度が改善した。

実験結果の考察のために、TFIDF 値を使った方法で重要索引語と判別された索引語のうち、TFIDF 値の上位 10 語を表 5.3.1.1 に示す。全ての分野で、該当分野の特徴を表していない索引語を重要索引語と見なしてしまっている。例えば、‘ 核－核抑止力 ’ の「メリット」は核抑止力を持つことによるメリットという文脈の中で使われているが、メリットという索引語は核抑止力だけを特徴づける索引語ではなく、重要索引語としてはふさわしくない。‘ 原子力－原発反対 ’ の「大阪府」、「豊中市」は、原発反対運動が大阪府や豊中市で行われ、この索引語が頻繁に記事中に現れたため、選ばれているが、原発反対を特徴づける索引語ではない。同様に、‘ 環境－地球温暖化 ’ の「カー」、「課題」、「家計簿」‘ 福祉－政策 ’ の「経済」、「要旨」、‘ 通信－無線 ’

の「兵庫県南部地震」、「局長」、‘情報－情報開示’の「扉を開く」[2]、「劇薬」などについても、それぞれの分野を特徴づける索引語ではない。しかし、‘核－核抑止力’の場合には「共同防衛構想」、‘福祉－政策’の場合には「老人福祉」、‘情報－情報開示’の場合は「市民団体」、「条例改正」など、各分野だけに完全に排他的に現れる索引語ではないにしても、各分野の特徴を比較的よく表している索引語が重要索引語として選ばれており、この 3 つの場合の f 値の改善に寄与していると考えられる。

次に、本実験で使用したキーワードと正解セットの組ごとに、本 TFIDF 値を使った方法の結果と、キーワードを 2 回以上含む全記事中の全索引語数と、正解セット中の全索引語数の割合の関係を、表 5.3.1.2 にまとめた。このキーワードを 2 回以上含む全記事は各々のクラスタリングの際にベースとなる文書集合であり、正解セットの文書はユーザからフィードバックされ、TFIDF 値を使った方法で重み付けされる索引語を含んでいる。したがって全記事中の全索引語数と正解セット中の全索引語数の割合は、クラスタのベースとなる文書集合にどのくらいの割合でフィードバックするかを表している。

キーワード－正解セット	TFIDF 法の結果	*1
核－核抑止力	70-90%で小幅に改善	202/5420(3.7%)
原子力－原発反対	変化無し	132/2870(4.5%)
環境－地球温暖化	変化無し	474/5209(9.1%)
福祉－政策	10-80%で小幅に改善	503/2671(18.8%)
通信－無線	変化無し	106/4348(2.4%)
情報－情報開示	20-50%で改善	826/9492(8.7%)

*1:正解セット中の全索引語数／キーワードを 2 回以上含む全記事中の全索引語数

表 5.3.1.2　キーワードと正解セットの特性と TFIDF 法の結果

TFIDF 値を使った方法は、重要な索引語を選び、その索引語の重みを増して精度や再現率の向上を期待する方法であるので、どの索引語を重要と判定するかと並ん

[2] この「扉を開く」は、例えば、‘そんな思いで情報公開企画「扉を開く」を５月下旬から始めました。’のように文中に括弧つきで現れるので、このままで索引語となっている。

で、どのくらいの数や割合の索引語の重みを増すかについても重要である。ここで、このフィードバックする割合と、TFIDF 法による結果を比較してみると、最も結果が改善された‘情報－情報開示’の場合、フィードバックされ重み付けされる索引語数が 826 個と最も多く、割合も 8.7%と‘福祉－政策’に次いで大きい。その‘福祉－政策’の場合も小幅ながら改善が見られ、フィードバックされ重み付けされる索引語数が 503 個で 2 番目に多く、割合は 18.8%で最も大きい。‘原子力－原発反対’、‘通信－無線’の場合はフィードバックされ重み付けされる索引語数がそれぞれ 132 個、106 個と少なく、全記事中の全索引語数と正解セット中の全索引語数の割合も、それぞれ 3.7%、2.4%と小さいため、フィードバックの効果が十分ではなく、結果も改善しなかったと考察される。‘環境－地球温暖化’の結果については、この考察の範囲では明確な因果関係はうかがえない。

核－核抑止力		原子力－原発反対		環境－地球温暖化	
索引語	TFIDF 値	索引語	TFIDF 値	索引語	TFIDF 値
メリット	0.544443737	大阪府	0.671704032	カー	0.707106781
共同防衛構想	0.534668382	豊中市	0.671704032	環境カー	0.707106781
防衛構想	0.534668382	大阪外国語大学	0.644858596	課題	0.633235409
短距離	0.511426199	選挙	0.615374315	家計簿	0.55756216
搭載可能	0.511426199	選挙戦	0.615374315	計算	0.55756216
傘	0.478005678	兵庫県	0.610875082	安全省	0.547271832
ミサイル	0.449017246	兵庫	0.610875082	環境安全省	0.547271832
欧州	0.43522674	住民	0.598772905	売り上げ	0.515218444
共同	0.431257219	福井	0.566335152	環境対策	0.502834018
搭載	0.412510348	北陸三県	0.566335152	都道府県	0.502834018
福祉－政策		通信－無線通信		情報－情報開示	
索引語	TFIDF 値	索引語	TFIDF 値	索引語	TFIDF 値
経済	0.866441037	兵庫県南部地震	0.754591425	扉を開く	0.826452995
兵庫県南部地震	0.801969652	局長	0.662973744	市民団体	0.619195342
児童	0.706931149	地震	0.656194927	劇薬	0.54879078
要旨	0.659155218	義士	0.633401366	社民党	0.542783758
省庁	0.624567702	中近世ターム	0.633401366	運営	0.529664032
老人福祉	0.600032382	地方	0.592817906	支障	0.529664032
地震	0.597364778	記者	0.586496349	条例改正	0.520788398
学校	0.585026687	パーソナルコンピューター	0.521049619	議会情報	0.520788398

光	0.583860357	**米軍基地**	0.509431326	**議員提案**	0.520788398
プラン	0.583860357	**沖縄**	0.509431326	**廃棄**	0.510433429

表 5.3.1.1　TFIDF 値を使った方法で判別された重要索引語（上位 10 語）

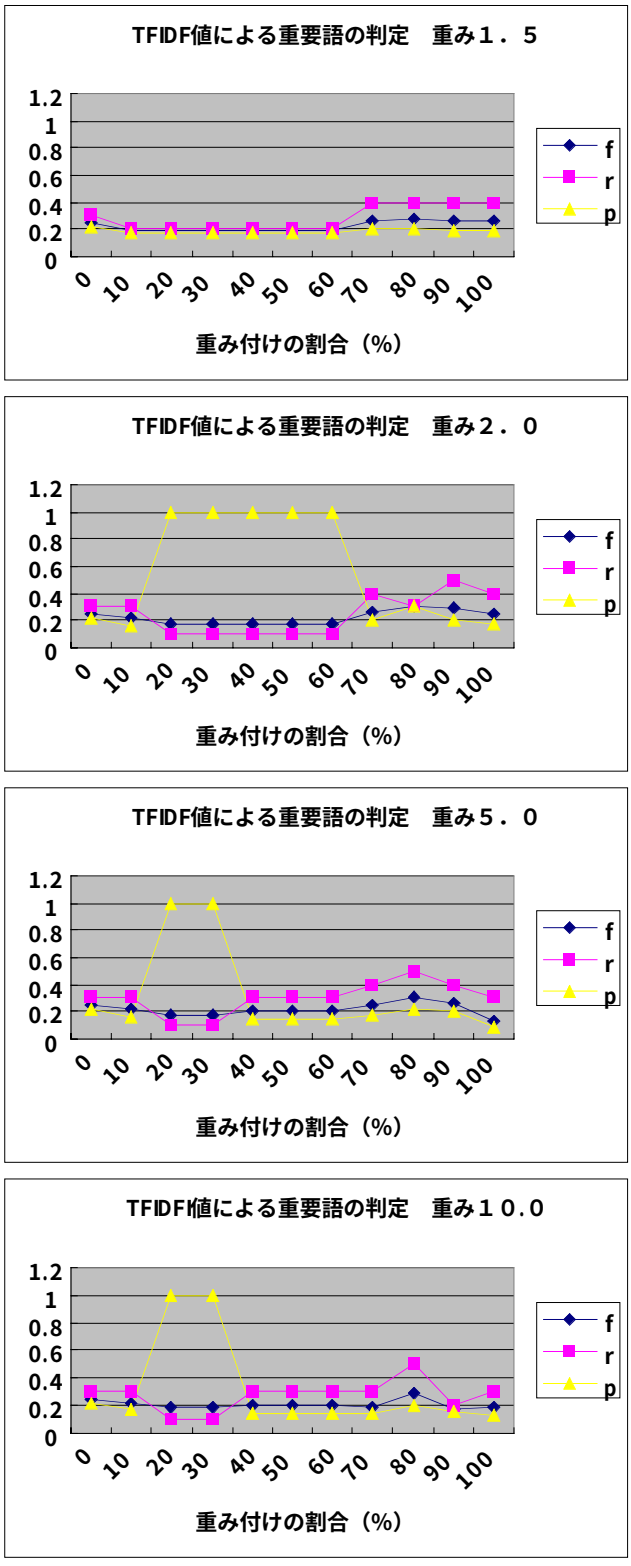
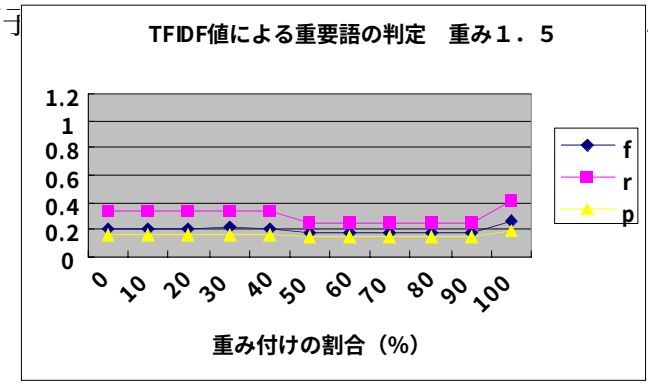
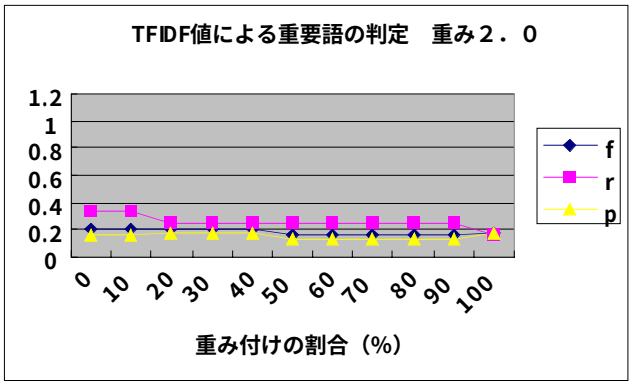
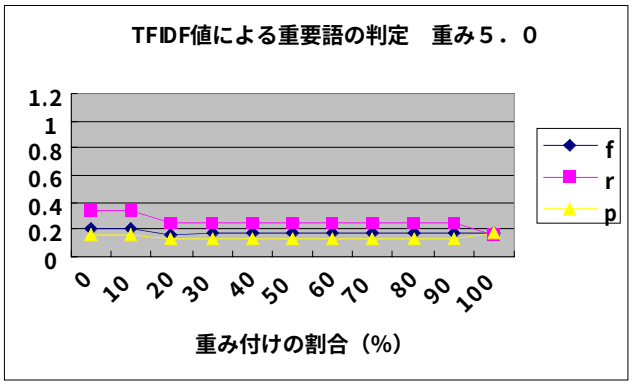


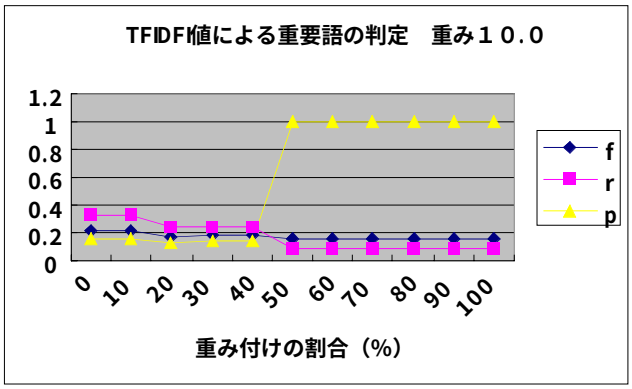
図 5.3.1.1　核ー核抑止力の TFIDF 法による適応クラスタリング結果

図 5.3.1.2　原子



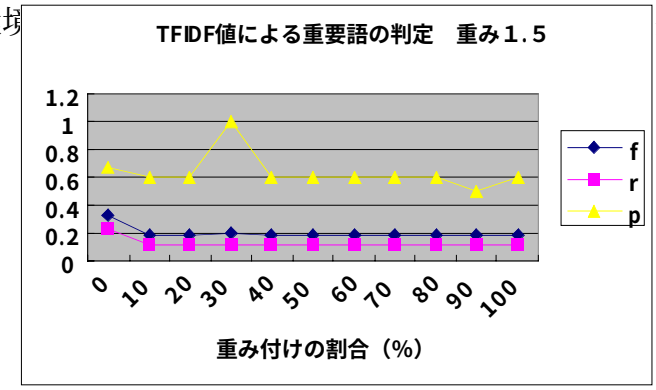


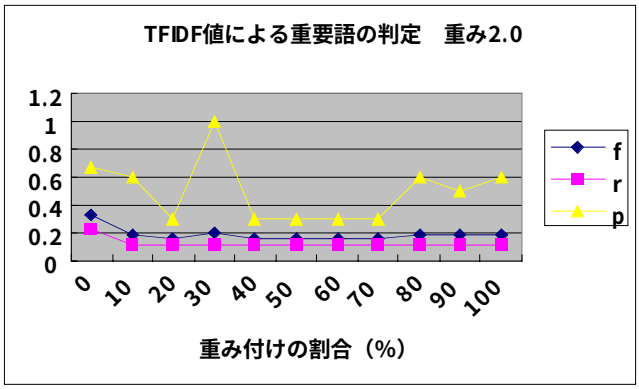


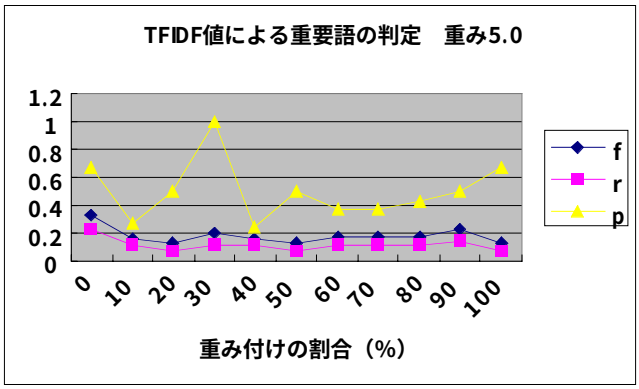


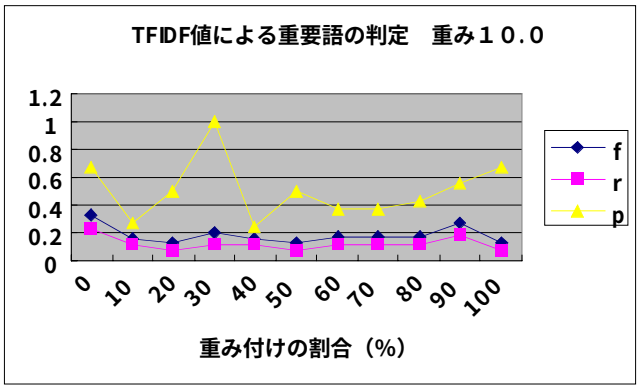
タリング結果

図 5.3.1.3　環境



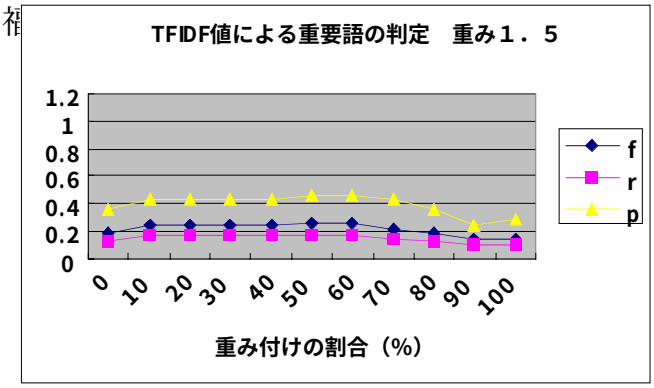


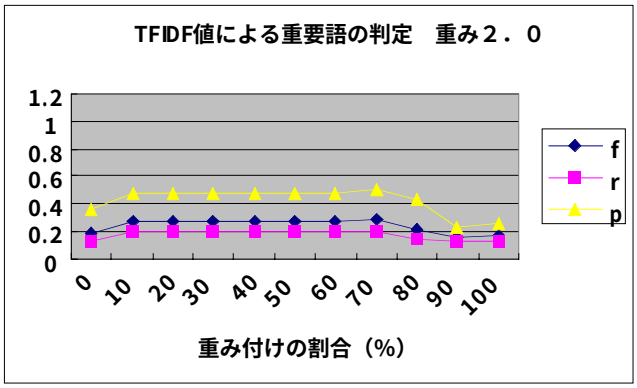


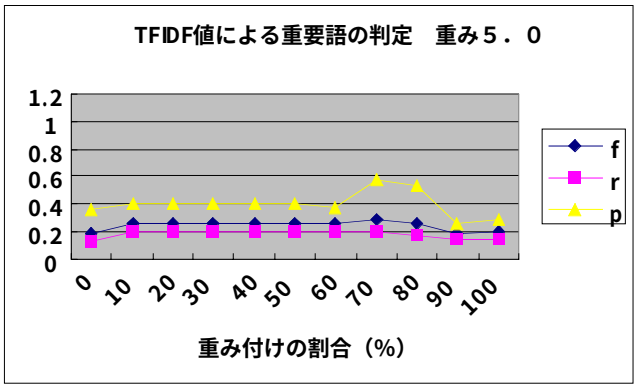


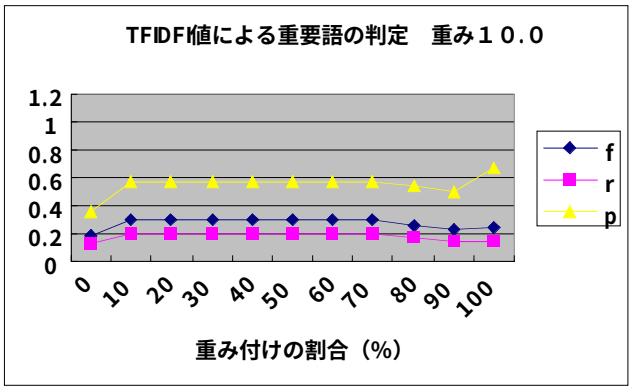
タリング結果

図 5.3.1.4　複









リング結果

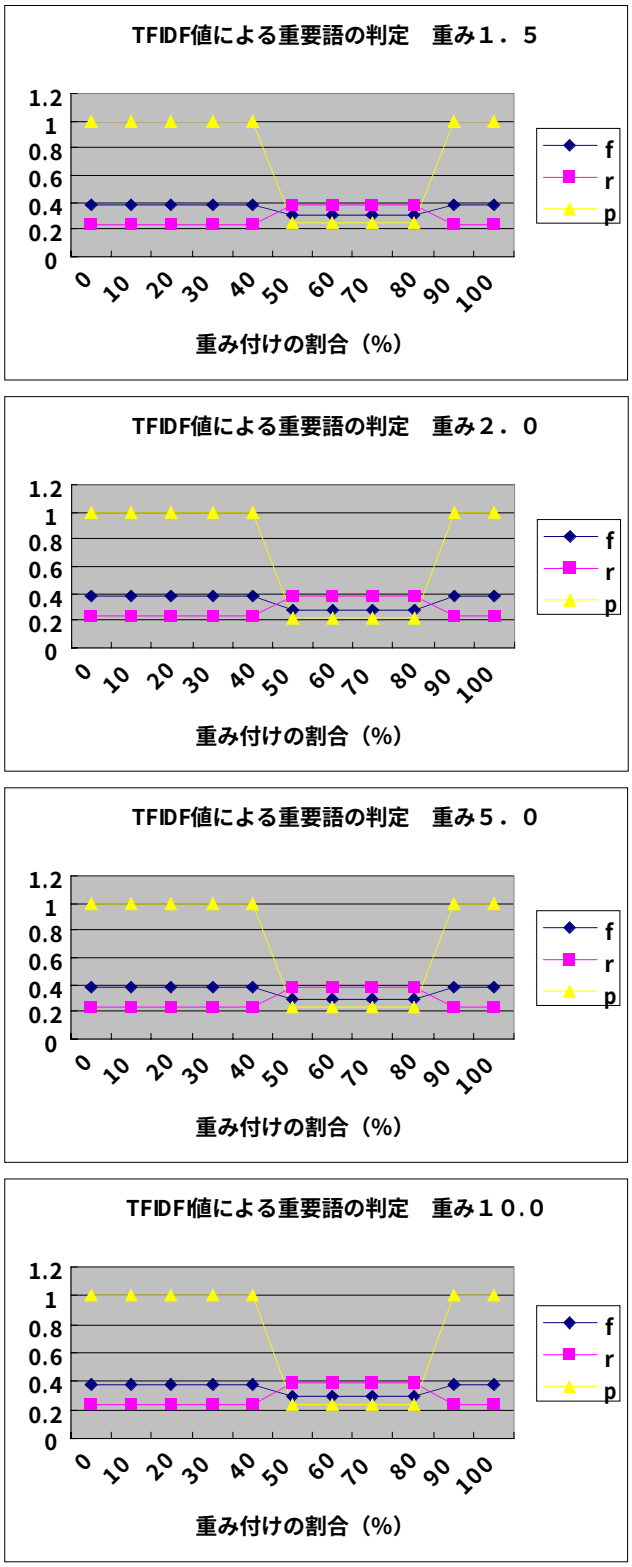


図 5.3.1.5　通信－無線通信の TFIDF 値を使った適応クラスタリング結果

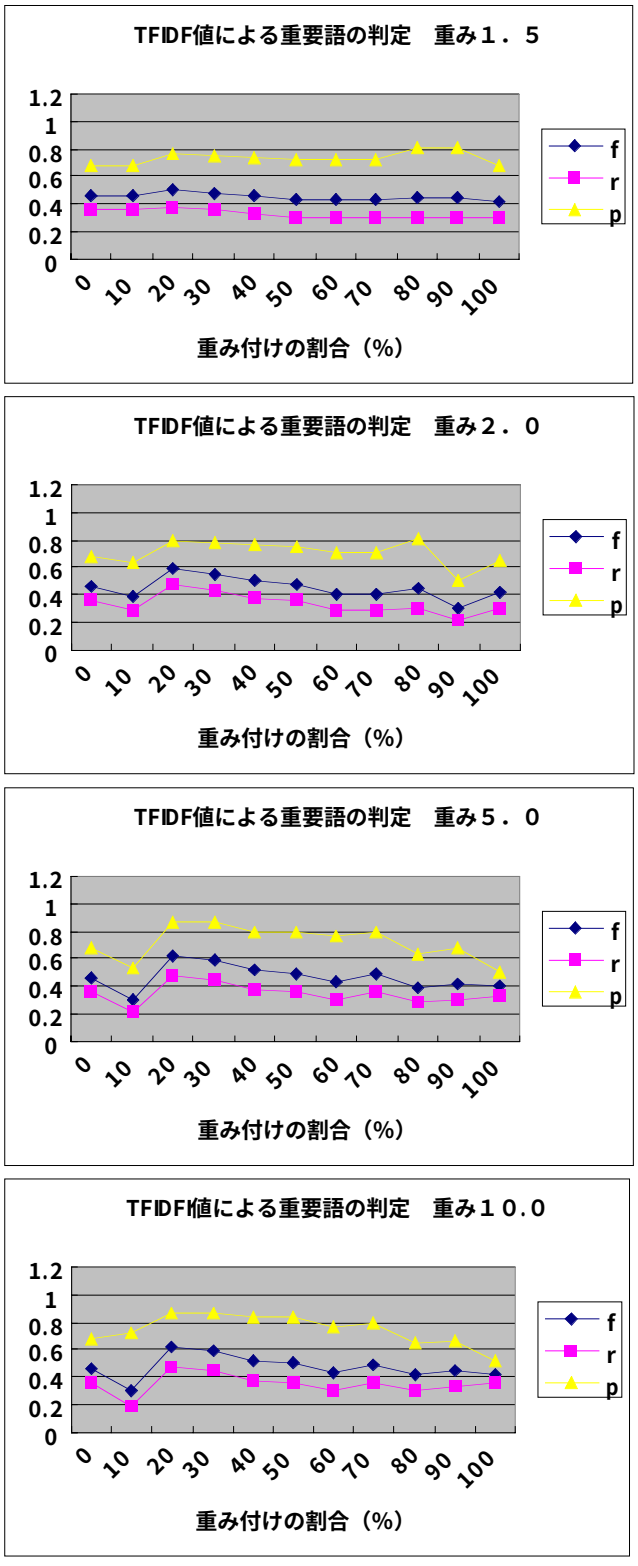


図 5.3.1.6　情報ー情報開示の TFIDF 値を使った適応クラスタリング結果

5.3.2　χ^2値を使った方法の結果と考察

実験結果を図 5.3.2.1～図 5.3.2.6 に示す。‘ 核－核抑止力 ’、‘ 福祉－政策 ’の場合は、重要索引語の割合を 10%～50%にとった時にｆ値の改善が見られた。‘ 原子力－原発反対 ’、‘ 通信－無線 ’の場合はｆ値の変化はほとんど無かった。‘ 環境－地球温暖化 ’の場合は、重みを 5 倍にとり重み付けの割合を 10%～70%の時、ｆ値の改善が見られた。‘ 情報－情報公開 ’の場合は重要索引語の割合が 10%～70%の時、ｆ値の改善が見られた。

ほとんどの場合、精度と再現率の両方がｆ値に改善に寄与している。重要索引語の重みとｆ値の関係については、‘ 核－核抑止力 ’、‘ 情報－情報公開 ’の場合は、重みを増すにつれｆ値のピークの値も増しているが、他の場合は明確な関係はない。

実験結果の考察のために、χ^2値を使った方法で重要索引語と判別された索引語のうちχ^2値の上位 10 語を表 5.3.2.1 にまとめた。この表によると、‘ 環境－地球温暖化 ’の「温暖化」、「ＣＯ２」、「温室効果ガス」、「温室効果」、‘ 情報－情報公開 ’の「情報公開」、「情報公開法」などこの２つの分野については、非常によく各分野の特徴を表す語を重要索引語としており、これがこの２つの分野でのｆ値の改善の原因と考えられる。しかし、‘ 通信－無線 ’の場合は、最もこの分野の特徴を表すと思われる「無線」、を重要と判断しているにもかかわらず、ｆ値が改善していないのは、初期クラスタの状態でもｆ値が 0.4 と比較的高く、精度に至っては１であり、これ以上のｆ値の改善を難しくしていると考えられる。‘ 原子力－原発反対 ’ではほとんど全ての重要索引語がこの分野とは無関係であり、結果的にｆ値もほとんど変化していない。全体的にχ^2値を使った方法は、‘ 環境－地球温暖化 ’のように非常に良く重要索引語を判定する場合と、‘ 原子力－原発反対 ’のように全く判定できていない場合があり、重要索引語抽出の精度に大きな差がある。

χ^2値を使った方法は、ある索引語のクラスタ内と文書内での出現頻度の偏りを計算し、その索引語が全文書に均等に分布している場合とのずれで、索引語のあるクラスタに対する重要度を判定する方法である。したがって、ある正解セット(＝クラスタ)にだけ偏って分布している索引語の存在が大前提になっている。この観点から今回実験に使った正解セットを調べてみると、‘ 環境－地球温暖化’の「温暖化」、「Ｃ

Ｏ２」、「温室効果ガス」、「温室効果」は環境分野の中でも地球温暖化に関する文書にだけ偏って出現する索引語であり、この事が、χ^2 値を使って精度良く重要索引語を判別できた理由であると考えられる。同じように、‘ 情報－情報公開 ’ の「情報公開」、「情報公開法」と通信－無線 ’ の「無線」についてもそれぞれの分野に特化して出現する索引語である。これに対して、‘ 原子力－原発反対 ’ の場合は、そもそも原発反対分野にだけ偏って分布している索引語が存在していないことが、重要索引語を判別できない理由ではないかと考える。

次に、前節と同じように、本実験で使用したキーワードと正解セットの組ごとに、本 χ^2 値を使った方法の結果と、キーワードを 2 回以上含む全記事中の全索引語数と、正解セット中の全索引語数の割合の関係を、表 5.3.2.2 にまとめた。χ^2 値を使った方法は、重要索引語の判定に χ^2 値を使っている点だけが、TFIDF 値を使った方法との違いであり、フィードバックされ重み付けされる索引語の数や割合による結果へ影響は全く同様に考察することが出来る。この表によると、TFIDF 値を使った方法と同様に、フィードバックされ重み付けされる索引語の数が少なく割合も小さい ‘ 原子力－原発反対 ’、‘ 通信－無線 ’ については、結果の改善が見られなかった。これに対し、フィードバックされ重み付けされる索引語の数が最も多い ‘ 情報－情報開示 ’、2 番目に多い ‘ 福祉－政策 ’ の場合は結果の改善が見られた。フィードバックされ重み付けされる索引語の数が比較的多い ‘ 環境－地球温暖化’、‘ 核―核抑止力 ’ の場合もある程度の改善が見られた。

キーワード－正解セット	χ^2法の結果	*1
核－核抑止力	10-50%で改善	202/5420(3.7%)
原子力－原発反対	変化無し	132/2870(4.5%)
環境－地球温暖化	10-70%(5 倍)で改善	474/5209(9.1%)
福祉－政策	10-50%で改善	503/2671(18.8%)
通信－無線	変化無し	106/4348(2.4%)
情報－情報開示	10-70%で改善	826/9492(8.7%)

*1:正解セット中の全索引語数／キーワードを 2 回以上含む全記事中の全索引語数

表 5.3.2.2　キーワードと正解セットの特性とχ^2法の結果

核－核抑止力		原子力－原発反対		環境－地球温暖化	
索引語	χ^2値	索引語	χ^2値	索引語	χ^2値
ドイツ	156.095466	議員	39.57683455	温暖化	108.965374
返還	99.47581016	兵庫県	39.57683455	ＣＯ２	59.77003091
沖縄	99.47581016	だまし	39.57683455	ガス	54.48268698
搭載	63.77477135	法案	39.57683455	温室効果ガス	54.48268698
フランス	62.87573178	選挙戦	39.57683455	温室効果	54.48268698
研究員	49.73790508	子供	39.57683455	対策	43.70236411
非常時	49.73790508	子供だまし	39.57683455	削減	43.70236411
サトゥ・リメイエ	49.73790508	運動	39.57683455	抑制	36.32179132
日記	49.73790508	プリヨンゴ	39.57683455	排出量	36.32179132
ドイツ領内	49.73790508	二酸化炭素	39.57683455	温暖化防止	36.32179132
福祉－政策		通信－無線通信		情報－情報開示	
索引語	χ^2値	索引語	χ^2値	索引語	χ^2値
厚生省	23.45153016	無線	322.1164477	情報公開	341.7025603
省庁	20.56603121	米軍	234.2594817	公開	277.9943003
中間報告	13.99383259	楚辺通信所	156.1729878	情報公開法	153.2567372
障害者福祉	13.99383259	米軍基地	156.1729878	接待	95.78546073
報告	13.99383259	沖縄	156.1729878	要綱	76.62836858
虐待	13.99383259	ヨット	78.0864939	公開法	76.62836858
見直し	13.99383259	発明	78.0864939	要綱案	76.62836858
学術会議	10.49537444	義士	78.0864939	訴訟	76.62836858
負担	8.999690811	救助	78.0864939	開示	64.15525837
消費税率	7.290284997	無線通信発明	78.0864939	原発	57.47127644

表 5.3.2.1　χ^2 値を使った方法で判別された重要索引語（上位 10 語）

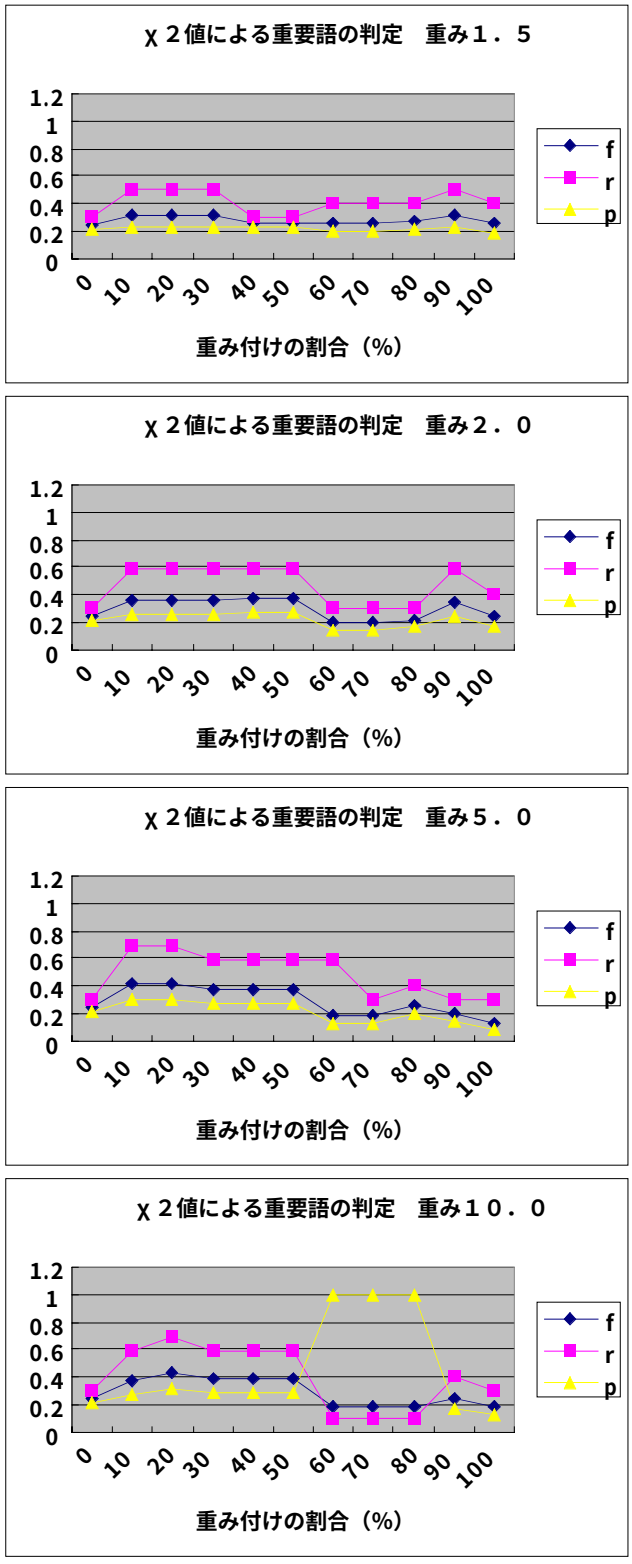


図 5.3.2.1　核ー核抑止力の χ^2 法による適応クラスタリング結果

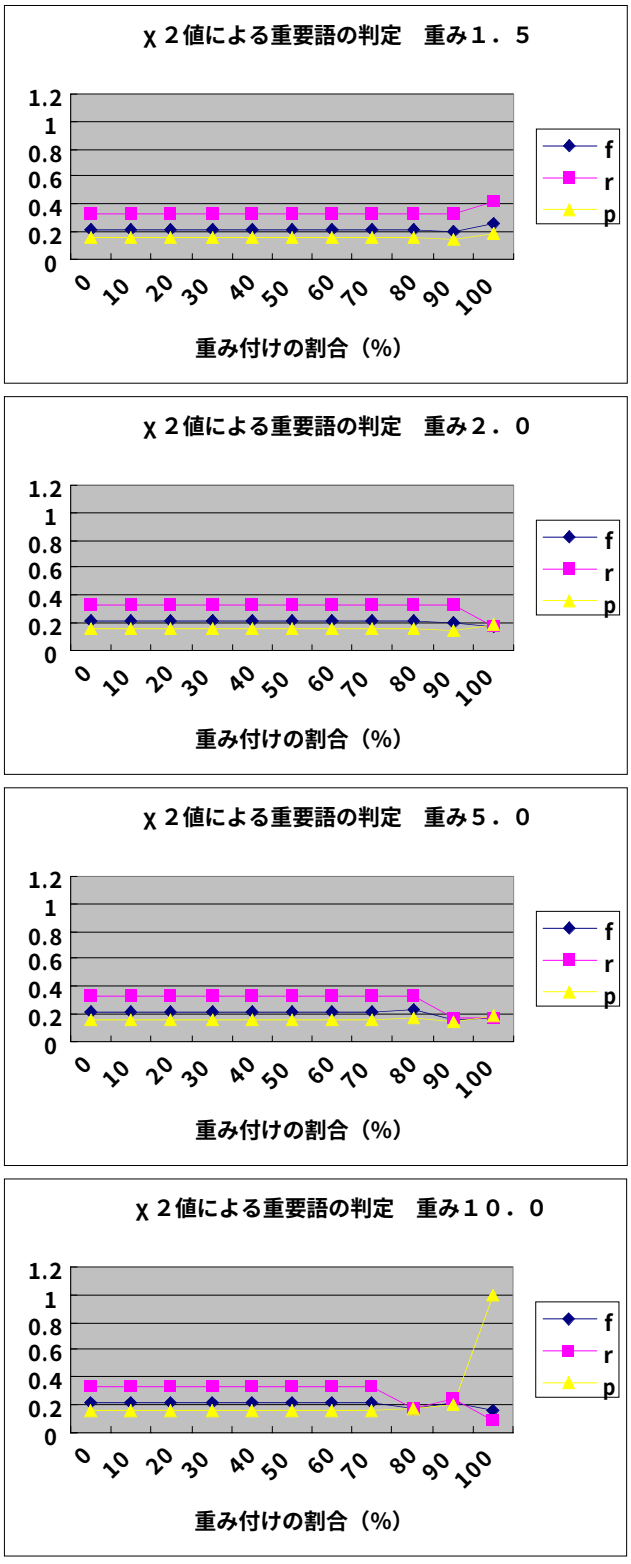


図 5.3.2.2　原子力ー原発反対の χ^2 法による適応クラスタリング

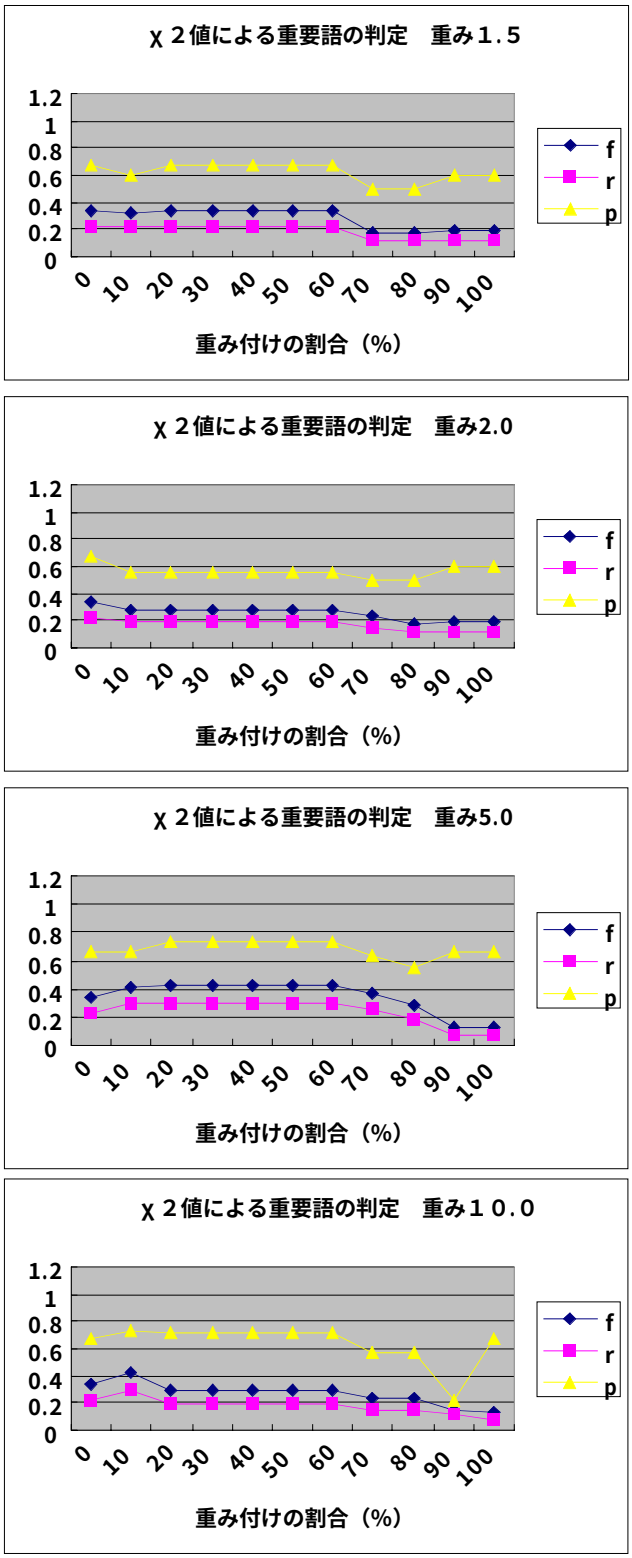


図 5.3.2.3　環境－地球温暖化の χ^2 法による適応クラスタリング

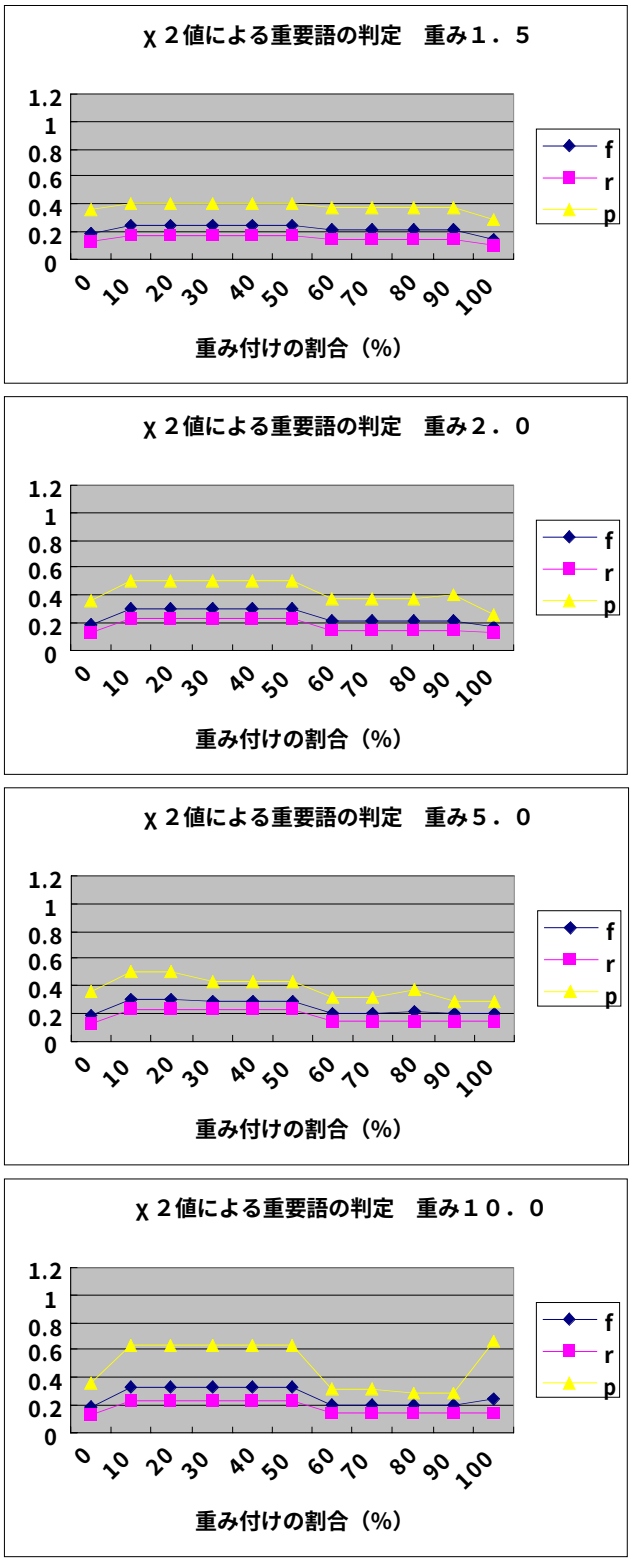


図 5.3.2.4　福祉－政策の χ^2 法による適応クラスタリング

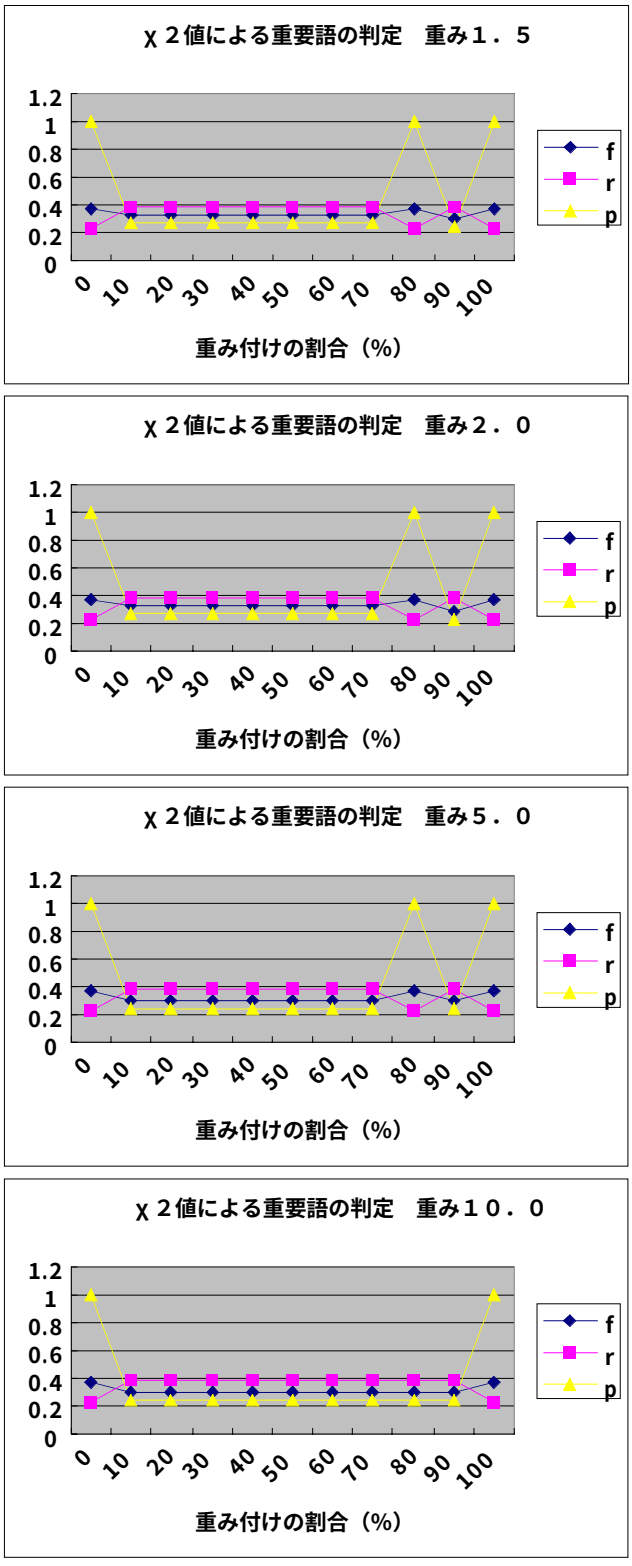
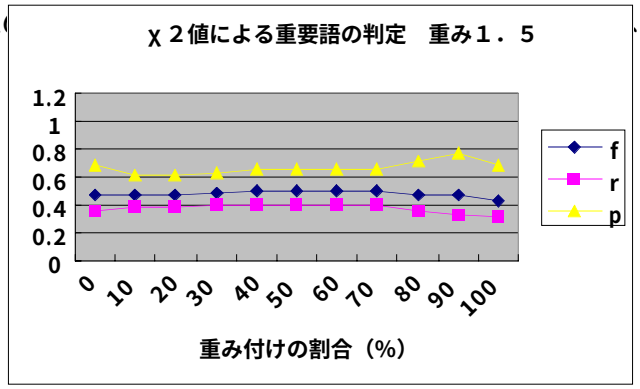
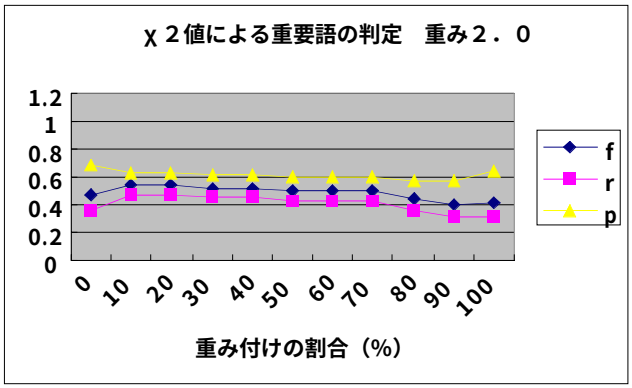
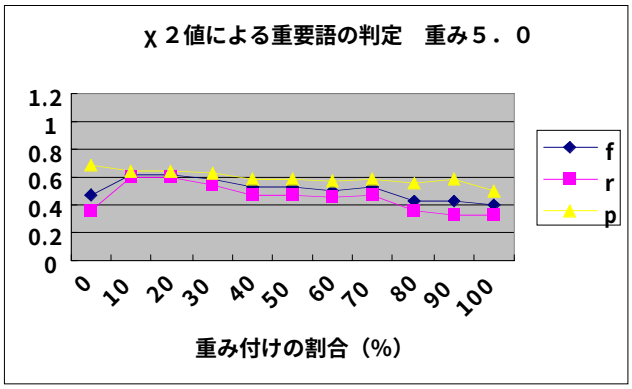


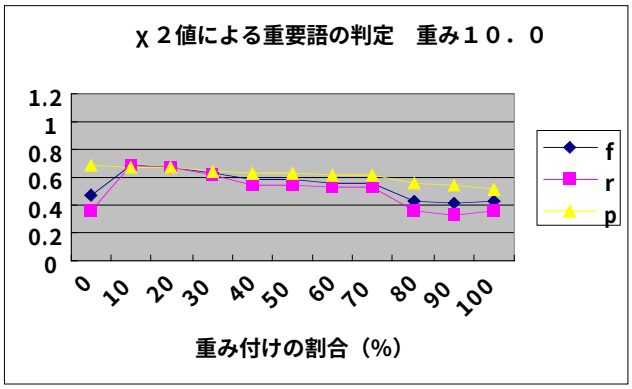
図 5.3.2.5　通信－無線の χ^2 法による適応クラスタリング

図 5.3.2.(..タリング


χ２値による重要語の判定　重み１．５
1.2
1
0.8
0.6
0.4
0.2
0
0 10 20 30 40 50 60 70 80 90 100
重み付けの割合（%）
f
r
p


χ２値による重要語の判定　重み２．０
1.2
1
0.8
0.6
0.4
0.2
0
0 10 20 30 40 50 60 70 80 90 100
重み付けの割合（%）
f
r
p


χ２値による重要語の判定　重み５．０
1.2
1
0.8
0.6
0.4
0.2
0
0 10 20 30 40 50 60 70 80 90 100
重み付けの割合（%）
f
r
p


χ２値による重要語の判定　重み１０．０
1.2
1
0.8
0.6
0.4
0.2
0
0 10 20 30 40 50 60 70 80 90 100
重み付けの割合（%）
f
r
p

5.3.3　文脈依存の度合いを使った方法の結果と考察

実験結果を、図 5.3.3.1～図 5.3.3.6 に示す。文脈依存の度合いを使った方法では f 値は小幅な変化にとどまっている。

本来この方法は、正しく分類されたトレーニングデータによる重要語と一般語の判別と、重み付けの学習を前提としているが、本研究では、あらかじめ正しく分類されたトレーニングデータは存在していないため、学習させることは出来ない。ここでは、重要語の判別式のみを利用した。

実験結果の考察のため、この方法で判別された重要索引語を表 5.3.3 に示す。この表によると、それぞれの分野に特徴的な語ではなく、各々のキーワードに関連している一般的な索引語を重要索引語と見なしている。例えば、‘ 核－核抑止力 ’ の「核開発」と「核爆弾」などは、「核抑止力」の特徴を表している索引語ではなく、「核」一般の特徴を表している索引語である。これが、ｆ値の改善につながらない原因と考えられる。他にも ‘ 原子力－原発反対 ’ の「原子力資料」、「原子力発電所」、‘ 福祉－政策 ’ の「介護」、「医療」などがこの場合にあてはまる。

文脈依存の度合いを使った方法とは、ある索引語が特定のクラスタに依存する度合いと、そのクラスタ内で特定の文書に依存する度合いを比較し、クラスタに依存する度合いの方が高い索引語を重要とする方法である。この依存する度合いは前節の χ^2 値の分散値で計算している。したがって、前節で考察した χ^2 値を使った方法が上手く機能する／しない要件は、文脈依存の度合いを使った方法でも同様に適用することが出来る。χ^2 値を使った方法は、ある正解セット(＝クラスタ)にだけ偏って分布している索引語の存在が大前提になっているため、偏って分布している索引語が存在しない場合は、全ての索引語の χ^2 値は比較的小さい同じような値になる。よって、文脈依存の度合いを使った方法は上手く機能しないことになる。今回の実験では、‘ 原子力－原発反対 ’ の場合に原発反対分野にだけ偏って分布している索引語がそもそも存在していないことが、文脈依存の度合いを使った方法で、重要索引語を判別できない原因であると考えられる。

しかし、χ^2 値を使った方法で重要索引語を判別できた、‘ 環境－地球温暖化 ’ の「温暖化」、「温室効果ガス」、「温室効果」や ‘ 情報－情報公開 ’ の「情報公開」、「情

報公開法」や通信－無線’の「無線」については、文脈依存の度合いを使った方法では、判別できていない。この原因としては、これらの索引語が、クラスタ間では分散が大きく、特定のクラスタに偏って出現しているにも関らず、そのクラスタ内の文書間でも分散が大きくなっているためと考えられる。例えば、‘環境－地球温暖化’の「温暖化」、「温室効果ガス」、「温室効果」などは、地球温暖化クラスタ内の文書の全てに必ずしも出現するわけではなく、むしろどれか 1 つだけの場合が多い。よって、結果的にそれぞれの索引語の文書内の分散が大きくなり、選ばれなかったと考えられる。

核－核抑止力		原子力－原発反対		環境－地球温暖化	
会談	大国	電力		日本	審議会
核開発	再開	原子力資料		中国	安全省
米国	フランス	数詞		問題	共生
問題		資料		世界	削減
政策		原子力発電所		シンポジウム	首相
大統領		原子力立地		白書	環境白書
核爆弾		情報室		省庁	地球環境
外交		原発		中央環境審議会	環境ＮＧＯ
開発		みんなの広場		対策	目標
協議		島根		配慮	ＣＯ２
配備		原子力資料情報室		日独	米
核実験再開		発電施設		都市	アンケート
核実験		原子力		ＮＧＯ	環境審議会
爆弾		資料情報室		生活	環境安全省
ボタン				案	防止
核				選定	経済
年間				対立	
福祉－政策		通信－無線通信		情報－情報開示	
介護	コスト	パーソナルコンピューター		Ｇ７	市民
政治	サービス	パソコン		報告	要望書
医療	年代	コンピューター		住民	グループ
負担	大阪府	大学		交換	後退
国会	高年	地方		開示	計画
年金	家族	マルチメディア		単独	信用
法案	障害者	数詞		都	取引
非分類語	ビジョン	通信		検査	自治体
高齢者				部門	事故
老人				改正	中間報告
充実				通産省	情報公開
数詞				公開	がん
福祉ビジョン					
				毎日新聞	予約
福祉				弁護士	情報
家庭				情報交換	テロ
生活				情報開示	
予算				義務	

表 5.3.3　文脈依存の度合いを使った方法で判別された重要索引語

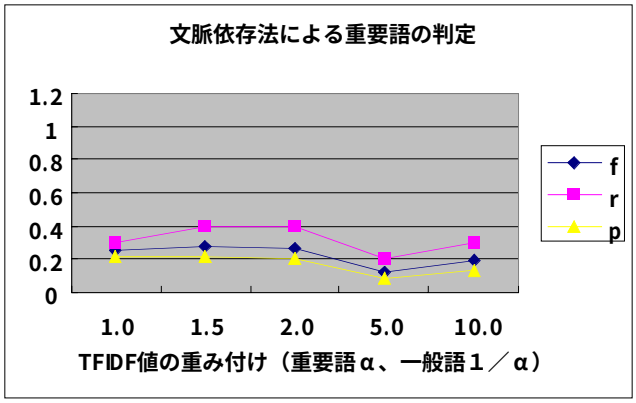


図 5.3.3.1　核－核抑止力の文脈依存の度合いによる適応クラスタリング

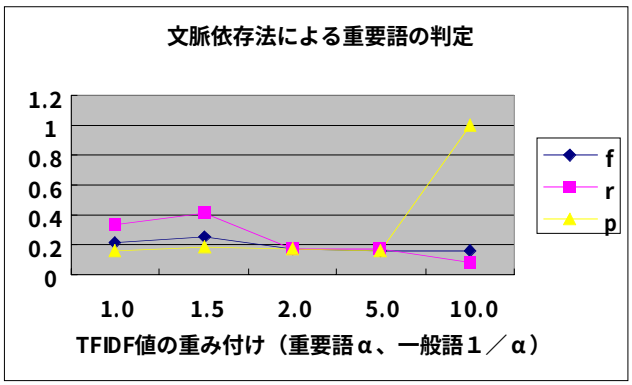


図 5.3.3.2　原子力－原発反対の文脈依存の度合いによる適応クラスタリング

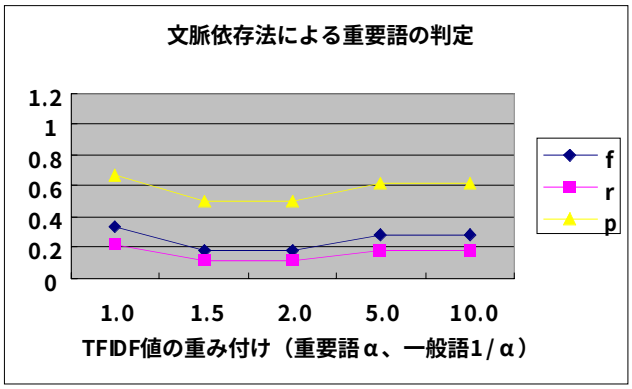


図 5.3.3.3　環境－地球温暖化の文脈依存の度合いによる適応クラスタリング

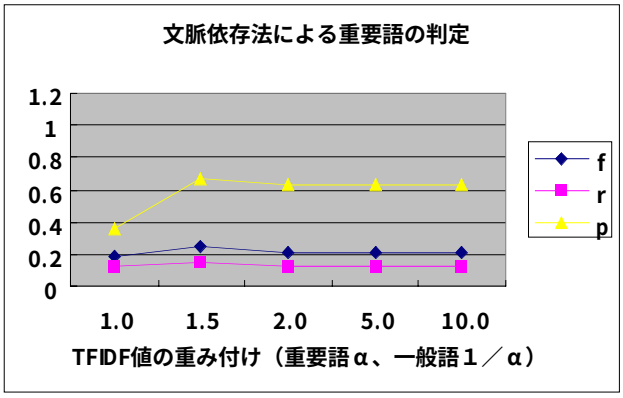


図 5.3.3.4　福祉－政策の文脈依存の度合いによる適応クラスタリング

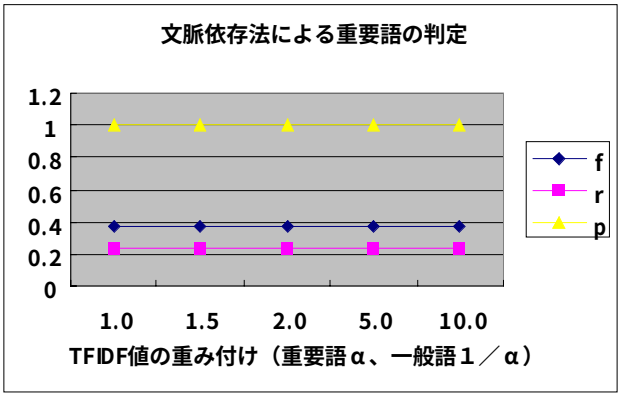


図 5.3.3.5　通信－無線の文脈依存の度合いによる適応クラスタリング

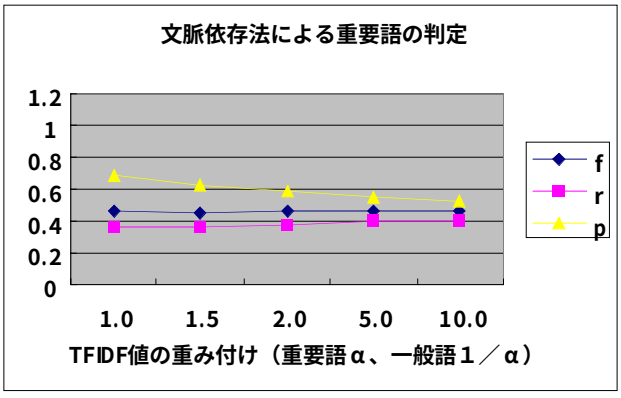


図 5.3.3.6　情報－情報開示の文脈依存の度合いによる

5.3.4　見なし共起を使った方法の結果と考察

実験結果を、図 5.3.4.1～図 5.3.4.6 に示す。見なし共起を使った適応クラスタリングにより、いずれのキーワードにおいてもｆ値の改善が見られた。なお、‘ 原子力－原発反対 ’ の重み 5 倍、10 倍と、‘ 通信－無線 ’ において精度が極端に変化し 1 になる箇所があるが、これはこの 2 分野の正解セットの大きさが小さいため、全くごみの無い小さなクラスタが最も ｆ 値の良いクラスタに選ばれてしまうためであり実質的な意味は無い。重要索引語の重み付けの大きさは、結果にほとんど影響しない。これは、見なし共起の結果生じる共起語の拡張による文書間の類似度の変化が、索引語の重み付けの変更の効果より相対的に大きいためであると考えられる。

全体的に、見なし共起の割合が低い間は、精度を保ちながら再現率が改善するためｆ値は改善するが、見なし共起の割合を高くするにしたがって、精度が悪化するため f 値も悪化する。このうち特に精度に関する結果を表 5.3.4 にまとめた。

キーワード－ 正解セット	見なし共起の割合と精度の推移	*1
核－核抑止力	見なし共起の割合が 60%までは精度は若干改善されるが、60%以上では悪化する	142/5420(2.6%)
原子力－原発反対	割合が 10%～90％までの間は、若干改善された精度を保つ	86/2870(3.0%)
環境－地球温暖化	割合が 10%で精度が落ち、その後若干改善し、50％から再び悪化する	278/5209(5.3%)
福祉－政策	割合が 10%～80％までの間は、若干改善された精度を保つがその後悪化する	299/2671(11.2%)
通信－無線	割合が 40%までは精度は悪化し、その後 80%までは比較的精度を保ち、その後再び悪化する	52/4348(1.2%)
情報－情報開示	割合が 10％の時、若干精度は改善するがその後は一様に悪化する	400/9496(4.2%)

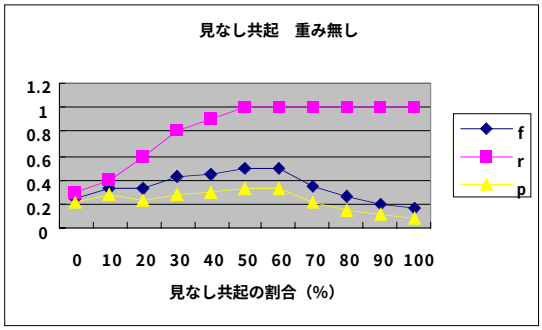
*1:正解セット中の共起しない索引語数／キーワードが 2 回以上現れる全記事中の全索引語数

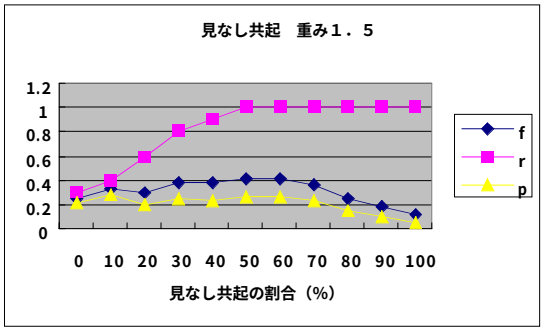
表 5.3.4　キーワードと正解セットの特性と見なし共起法の結果

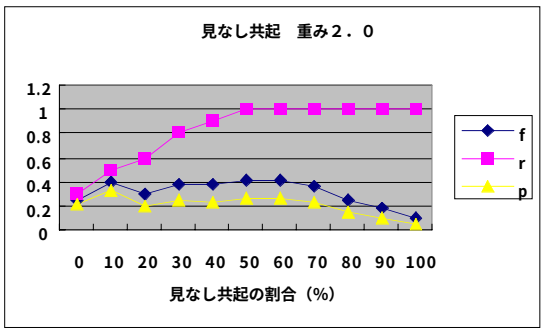
見なし共起を使った方法では、正解セット中の索引語のうち共起しない索引語を共起したと見なす。よって、正解セットの文書中の索引語は全て共起したことになり、再現率が改善されるのは自明である。その反面、本来なら全く関係ない文書を集めてしまう効果も当然ながらあり、精度は悪化することが容易に想像される。しかし、現実には精度が若干ながら改善したり、ほとんど変化しない例も見られる。

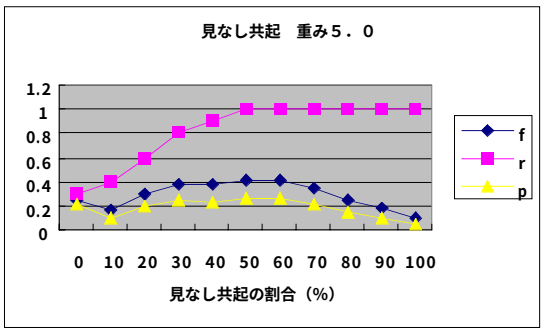
見なし共起の効果を考察する上で、正解セット中の共起しない索引語（＝見なし共起させる索引語）がクラスタリングの対象となる全記事中の全索引語中でどのくらいの割合で含まれているかが重要な特性となる。表 5.3.4 によると、見なし共起させる索引語の割合が‘核－核抑止力’の場合で 2.6%、‘原子力－原発反対’の場合で 3.0%、‘通信－無線’の場合で 1.2%といずれも小さく、見なし共起の影響が比較的少ないと考えられる。実際、この 3 つの分野では、見なし共起させる索引語の割合が少ない間は、精度はおおむね悪化しない。しかし、見なし共起させる索引語の割合が大きい‘福祉－政策’、‘環境－地球温暖化’、‘情報－情報開示’でも精度が悪化しない場合があり、見なし共起させる索引語の割合と精度との関連は必ずしも明確ではない。

さらに、精度と f 値に関する考察で重要な点は、初期クラスタの状態で精度が高い場合と低い場合で、見なし共起させる索引語の割合を大きくしていった際の、f 値と精度の変化の仕方が違う点である。精度が初期クラスタ状態ですでに高い場合は、再現率の上昇が精度の下降を補償しきれない場合に f 値が悪化する。一方、精度が初期クラスタの状態で低い場合は、見なし共起の割合を高くしていくにつれて、最初は再現率が上昇し、f 値も小幅な改善が見られる。しかし、見なし共起の割合がある程度高くなると、再現率の上がり方よりも精度の下がり方の方が大きくなり、ｆ値も劣化する。









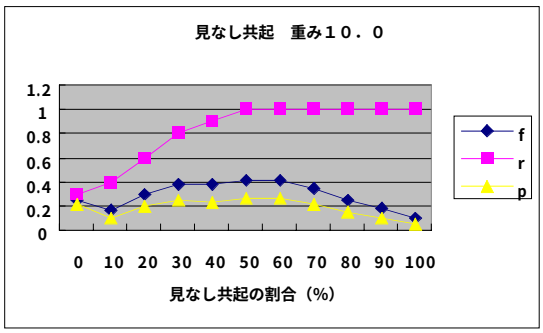
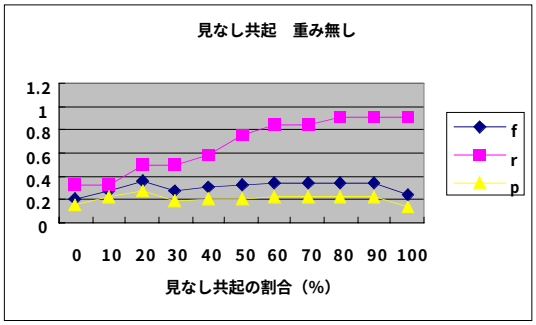
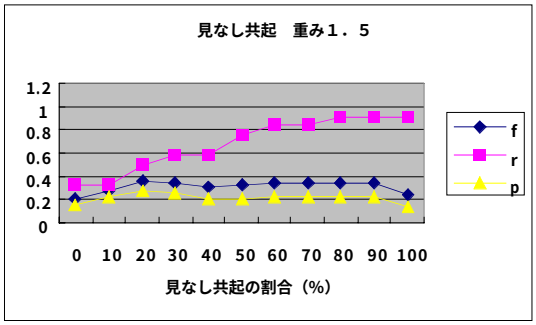
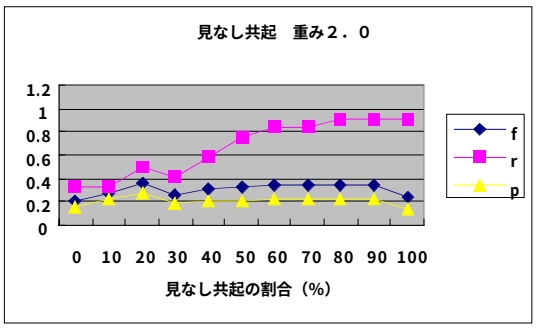
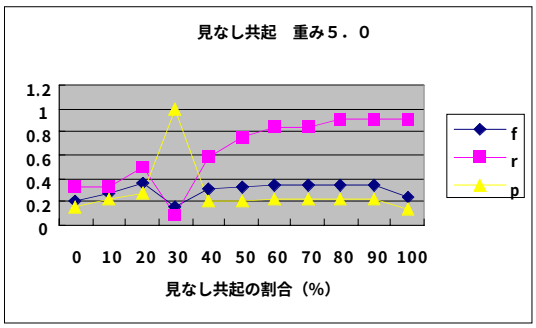


図 5.3.4.1　核ー核抑止力の見なし共起による適応クラスタリング結果









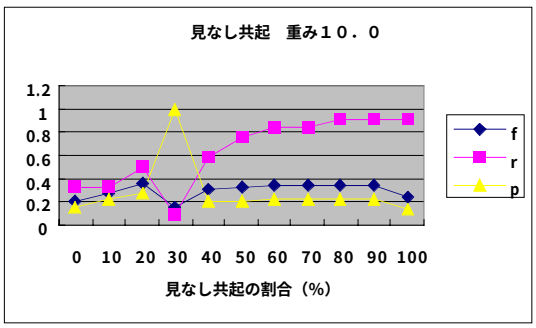
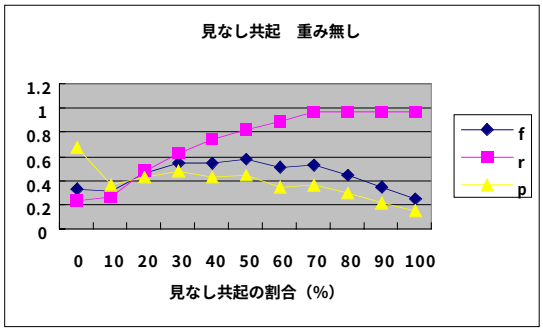
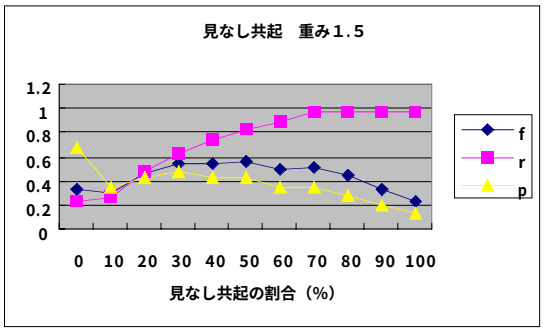
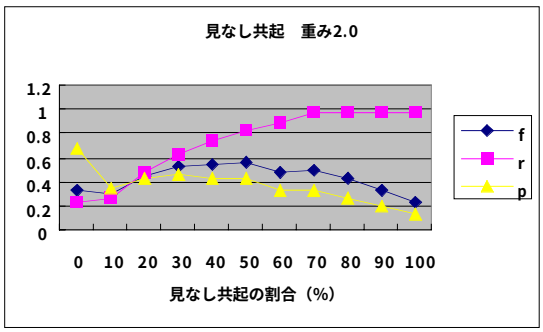
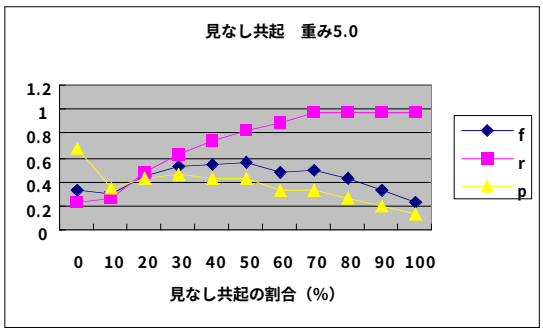


図 5.3.4.2　原子力一原発反対の見なし共起による適応クラスタリング結果









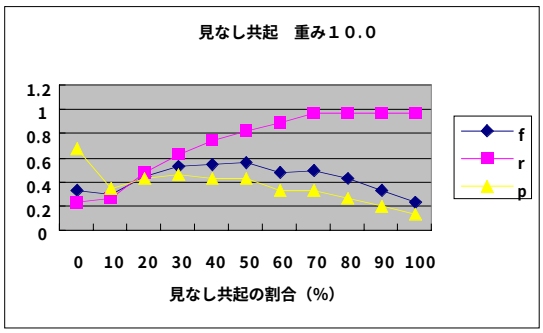
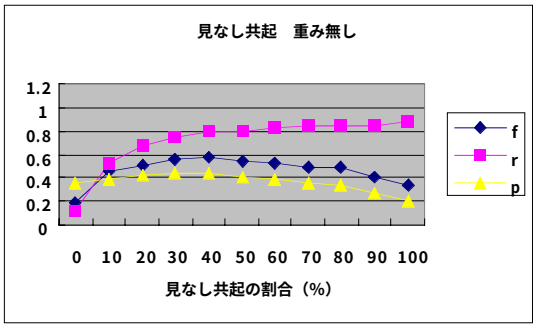
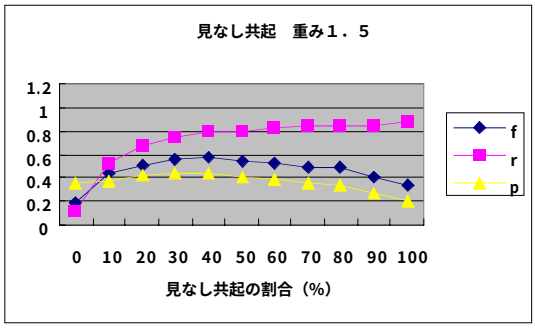
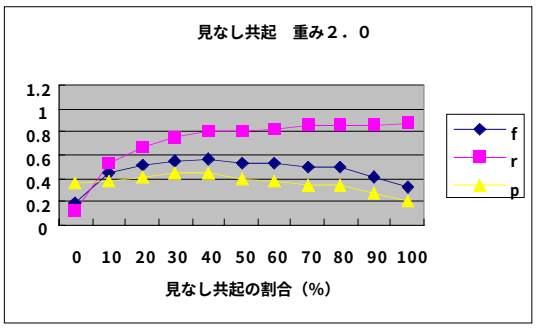
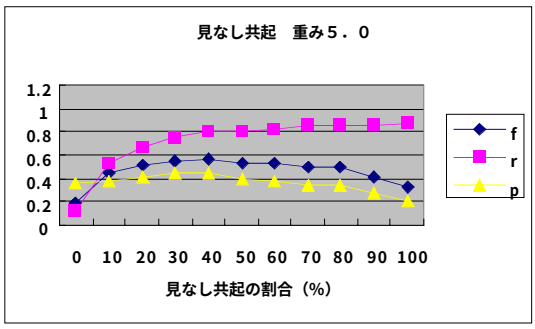


図 5.3.4.3　環境－地球温暖化の見なし共起による適応クラスタリング結果









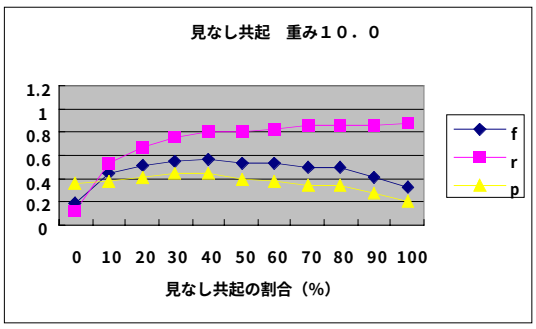
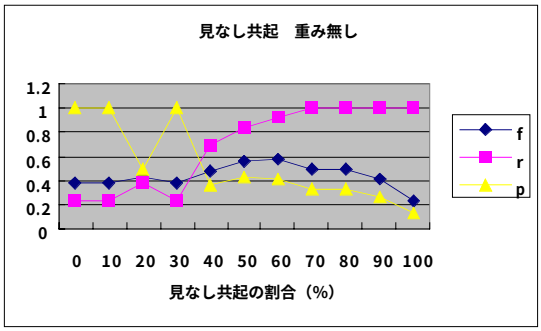
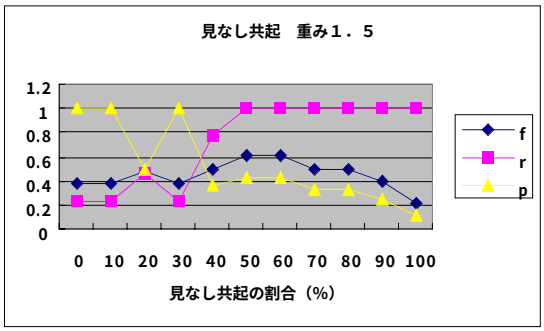
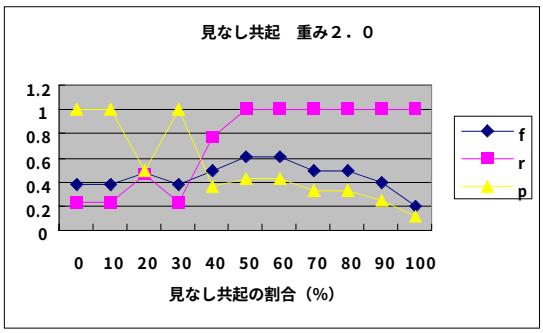
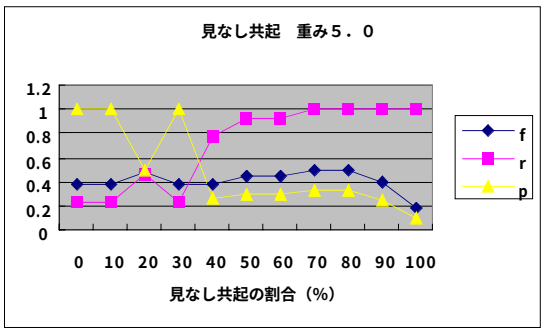


図 5.3.4.4　福祉一政策の見なし共起による適応クラスタリング結果









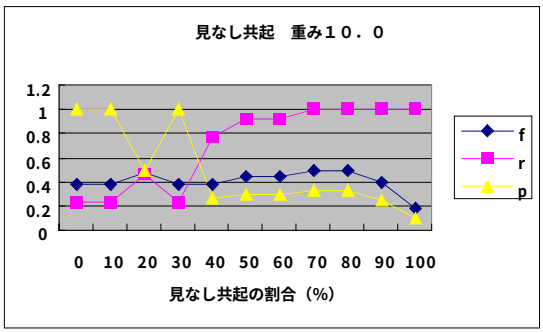
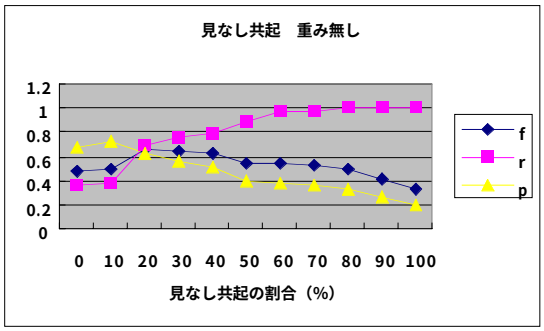
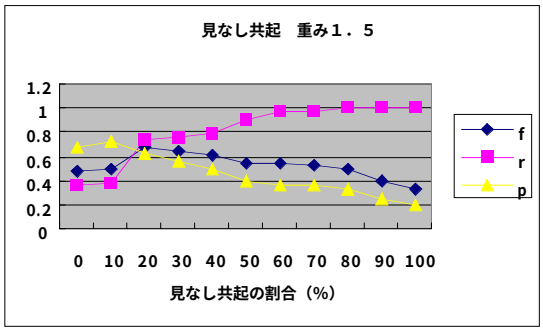
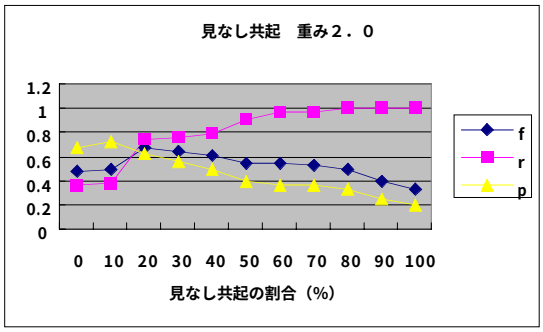
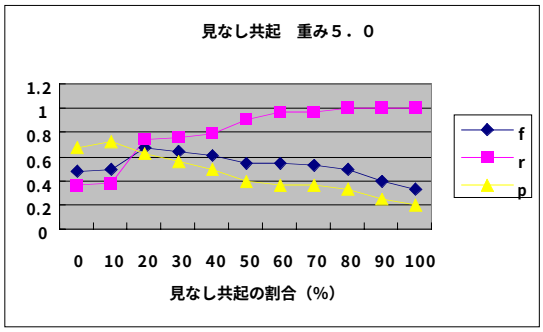


図 5.3.4.5　通信ー無線の見なし共起による適応クラスタリング結果









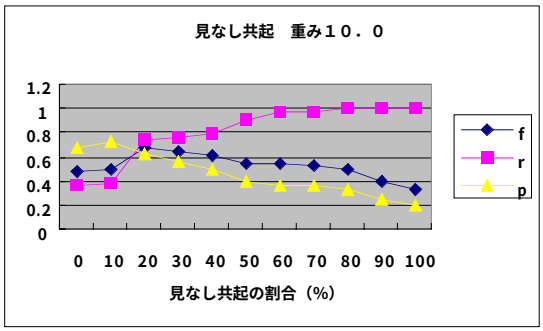


図 5.3.4.6　情報－情報公開の見なし共起による適応クラスタリング結果

5.3.5　考察のまとめ

この 4 つ方法の中では、TFIDF 値を使った方法が、精度を最もよく改善する方法である事がわかった。重要索引語の見なす割合と重みの決定は、TFIDF 値を使った方法で、キーワードごとに、最も f 値が大きくなる割合と重みを平均することにより計算する。表 5.3.5.1 に f 値が最大になる割合と重みを示す。

	f 値が最大になる割合(%)	f 値が最大になる重み
核－核抑止力	80	5.0
福祉－政策	70	5.0
情報－情報公開	20	5.0
（平均）	56.7	5.0

表 5.3.5.1　f 値が最大になる割合と重み

今回の実験の範囲では、TFIDF 値を使った適応クラスタリング方法で、重要索引語の割合を 56.7%、重みを 5 倍にした時に、f 値が最も良くなることがわかった。

見なし共起を使った適応クラスタリングは、ユーザからフィードバックされた記事を 1 つのクラスタにまとめる性能が高いことが確認された。重みと見なす割合の最適値を算出する。重み付けの大きさは結果にほとんど影響しないので、重み付け無しとする。重み付けがない場合に、最も f 値が大きくなる見なす割合を、表 5.3.5.2 に示す。

	f 値が最大になる割合（%）
核－核抑止力	60
原子力－原発反対	20
環境－地球温暖化	50
福祉－政策	40
通信－無線	60
情報－情報公開	20
（平均）	41.7

表 5.3.5.2　f 値が最高になる見なし共起の割合

今回の実験の範囲では、見なし共起による適応クラスタリングで重要索引語の割合を 41.7%、重み付け無しにした場合に、f 値が最も良くなることが分かった。

第６章
結論

6.1　本研究の成果

本研究では、クラスタリングによる社会情報の分類と時間軸を持つ空間への配置によって、ある概念の時間的変遷を可視化できるシステムを構築し、さらにクラスタリング結果に個人的な視点を反映させる方法である適応クラスタリングの様々な方法の性質を実験により検証した。

クラスタリングにより、ある概念を分類しリスト表示させると共に、クラスタ内の構造を時間軸を持った２次元上空間上にグラフ表示することにより時間的変遷を可視化することが出来た。

さらに、適応クラスタリングの評価実験により、TFIDF 値を使った適応クラスタリング方法が最も精度を改善し、索引語の割合を 56.7%、重みを 5 倍にした場合に、f 値が最高になることを明らかにした。また、見なし共起を使った適応クラスタリング方法が最も再現率を改善し、索引語の割合を 41.7%、重み無しにした場合に、f 値が最高となることが分かった。

6.2　今後の課題

今後の課題として検討する必要のある点を挙げる。

1. クラスタリングのアルゴリズムを改善し、クラスタリング速度を向上することにより、Web 上の情報のリアルタイムでの可視化を可能にする。関連研究でも述べたように、クラスタリングの精度を保ちながら線形時間で終了するアルゴリズムはすでにいくつか研究されている。これらのクラスタリングアルゴリズムを本研究に取り入れ、応答時間を改善する事を検討する。

2. 現在のシステムでは、1 つのクラスタ中の時間的変遷を可視化することが出来る。しかしながら興味のある問題の全体像をつかむためには、全クラスタの構造やクラスタ間の関係を比較したい場合が出てくる可能性がある。原理的には本システムの方式を全クラスタの可視化に適用することも可能ではあるが、文書数が多くなりすぎて表示が煩雑になる。クラスタ全体の時間的変遷をコンパクトに可視化できる方法を検討する。

3. 現在のシステムでは、ユーザはフィードバックにより 1 つの概念の整理をすることを前提としている。例えば、環境について「地球温暖化」と、「エコビジネス」の両方に興味のあるユーザは、本システムを使って各々の概念の整理を別々に行う必要がある。しかし、現実には「地球温暖化」が「エコビジネス」にどのような影響を与えているかなどを知りたい場合もある。このような場合は、複数の概念をまとめて整理するような機能が必要であり、今後の研究課題である。

謝辞

本研究を進めるにあたり、石崎雅人助教授には、数々のご指導･ご助言を頂き心から感謝いたします。

また、普段から様々な面でご指導頂いた Ho Tu Bao 教授、本研究と関連の深い副テーマにおいて的確なご指導を頂いた國藤進教授、本論文の中間審査、本審査において貴重なご指摘を頂いた中森義輝教授に深く感謝いたします。

最後に、普段の勉強会などで忌憚の無いアドヴァイスを頂いた石崎研究室の学生、研究生の皆様方に感謝いたします。特に、植田繁雄、中谷彰宏、松永政幸の各氏には評価実験の正解セット作成に協力していただきました。ここに感謝の意を込めて明記させて頂きます。

参 考 文 献

[Cutting 92] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey : Scatter/Gather: A Cluster-based Approarch to Browsing Large Document Collections, in Proceedings of the 15th Annual International ACM/SIGIR Conference, Copenhagen, 1992

[福本 99] 福本文代 鈴木良弥 : 語の重み付け学習を用いた文書の自動分類, 情報処理 Vol.40 No.4 1782-1791 1999

[Hearst 94] Marti A. Hearst : Using Categories to Provide Context for Full-Text Retrieval Results, In the Proceedings of the RIAO 94. 1994

[Hearst 95a] Marti A. Hearst : TileBars: Visualization of Term Distribution Information in Full Text Information Access, in Proceedings of CHI 95, CO, May 1995

[Hearst 95b] Marti A. Hearst, David R. Karger, Jan O. Pedersen : Scatter/Gatheras a Tool for the Navigation of Retrieval Results, In the Working Notes of the 1995 AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval. 1995

[Hearst 97] Marti A. Hearst, Chandu Karadi : Cat-aCone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy , in Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 97), 1997

[岸田 98] 岸田和明 : 情報検索の理論と技術 図書館･情報学シリーズ３ 剄草書房 1998

[清田 98] 清田陽司 黒橋禎夫 中村順一 長尾真 : 構文情報を利用した電子ニュース記事のクラスタリングシステムの作成と評価,自然言語処理 126-11 1998

[Luhn 57] Luhn, H.P. : A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1(4), 390-317, 1957.

[毎日新聞]　CD－毎日新聞 91～97 年版

[松本 99]　松本裕治　他．茶筌 version 2.02 + IPA 品詞体系日本語辞書 version 2.1、奈良先端科学技術大学院大学　松本研究室 1999

[三末 94]　三末和男,杉山公造 : マグネティック・スプリング・モデルによるグラフ描画法について, ヒューマンインタフェース　グループウェア,17-24 1994

[長尾 76]　長尾真 水谷 幹男 池田 浩之 : 日本語文献における重要語の自動抽出, 情報処理 Vol.17 No.2 1976

[Salton 88] G. Salton, C..Buckley : Term-weighting approaches in automatic retrieval. Information Processing & Management, 24(5):513-523, 1988.

[Shneiderman 00] Ben Shneiderman, David Feldman, Anne Rose : Visualizing Digital Library Search Results with Categorical and Hierarchial Axes, in Proceedings of the fifth ACM conference on ACM 2000 digital libraries　2000.

[杉山 93]　杉山公造 グラフ自動描画法とその応用、計測自動制御学会編 1993

[徳永 99]　徳永健伸 : 情報検索と言語処理 言語と計算５　東京大学出版会　1999

[豊浦 97]　豊浦潤 徳永健伸 井佐原均 岡隆一 : RWC における分類コードつきテキストデータベースの開発,信学技法 NLC96-13 (1996-07)

[渡辺 94]　渡辺靖彦 竹内雅人 村田真樹 長尾真: χ^2 法を用いた重要漢字の自動抽出と文献の自動分類, 電子情報通信学会研究会 言語理解とコミュニケーション 94-25 1994

[Zamir 98] O Zamir and O Etzioni. Web document clustering: A feasibility demonstration. In Proc of the 21th Intl ACM SIGIR Conf, pages 46--54, 1998.

[Zamir 99] Oren Zamir, Oren Etzioni : Grouper: A Dynamic Clustering Interface to Web Search Results, The Eighth International World Wide Web Conference　1999.