

ユーザの視点を考慮に入れた社会情報可視化システムの構築

岩本雅道†, 石崎雅人‡, 國藤進‡

† 国際協力事業団, ‡ 北陸先端科学技術大学院大学・知識科学研究科

† iwamoto@hn.vnn.vn, ‡ { masato, kuni }@jaist.ac.jp

要旨

本研究では、新聞などの大量の社会情報のユーザによる再構成を支援するために、1) 時間軸に沿った情報推移を可視化する機能、2) ユーザの視点を考慮して情報を再クラスタリングする機能を持った社会情報可視化システムを構築した。再クラスタリングについては、索引語の重みづけ手法、索引語の決定手法に関して実験を行い、共起索引語が少ない場合には、見なし共起を、それ以外の場合、再現率を改善したい場合は、見なし共起を、精度を改善したい場合は、TFIDF 値または χ^2 値を用いるのがよいことが示唆された。

Building A Visualization System for Social Information With Re-clustering

Masamichi Iwamoto*, Masato Ishizaki**, Susumu Kunifuji**

*Japan International Cooperation Agency

**School of Knowledge Science, Japan Advanced Institute of Science and
Technology

*E-mail: iwamoto@hn.vnn.vn

**Email: { masato, kuni }@jaist.ac.jp

Abstract

To support the user's re-structuring of a large amount of social information such as newspaper articles, we build a social information visualization system with the functions of 1) displaying the information using the terms, their spatial relations on time axis and the original text and 2) re-clustering the information based on the user's feedback. To realize effective re-clustering, the experiments concerning the term weighting and selection methods were conducted. The experimental results suggested that for the documents containing a small percentage of the co-occurring terms, the forced co-occurrence method is effective for improving the precision and recall. For the documents full of co-occurring terms, the TFIDF or χ^2 based method should be used for better recall and the forced co-occurrence for better precision.

1. はじめに

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 World Wide Web (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。

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再クラスタリングを行う。

2. 関連研究

2.1. クラスタリングを利用した情報可視化システム

Scatter/Gather システム[1, 3]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。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。

Grouper システム[9]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。テーブルの 1 つの行が 1 つのクラスタを表す。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、文書タイトルのサンプルで要約される。フレーズには、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が付与される。

2.2. 再クラスタリング手法

クラスタリング結果にユーザ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。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。

一方、Grouper システムでは、ユーザの視点を取り入れる方法として、メイン結果ページのクラスタのうち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。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、検索結果を絞込むことになる。

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、分野依存度に基づいて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[2]。重要な索引語の判定には、 χ^2 値を使うもの[5, 8]。

2.3. 本システムとの関連

本項では、前項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情報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。

- ・ 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。
- ・ 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。

3. 社会情報可視化システム

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したり、各クラスタの構造を時系列に沿って空間表示することも出来る。システムによって自動生成されたクラスタが、ユーザの分類と異なっている場合は、ユーザ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 (索引語

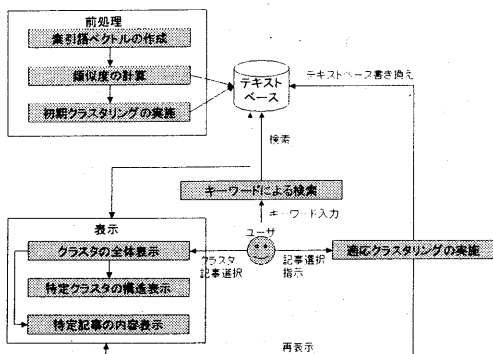


図 3.1 システムの概要

頻度-逆文書頻度 (TFTDF))、類似度の計算 (内積)、初期クラスタリング (郡間法) は、予め実施しておいた。今回、元文書群としては、CD-ROM 版の毎日新聞 (91 年~97 年) [4] を用いた。文字処理は Perl (1400 ステップ)、空間表示は Java 言語のアプレット (500 ステップ)、CGI は C 言語で実装した。

3.2. ユーザインタフェース

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。

- ・ 全体結果表示: ユーザ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。
- ・ 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。
- ・ 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。

3.2.1. キーワード選択画面

本システムを起動すると、最初に図 3.2 に示すようなキーワード選択画面になる。ユーザ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

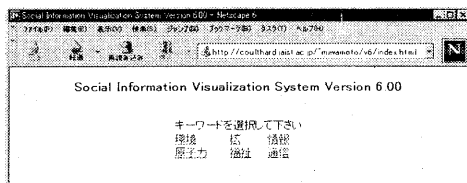


図 3.2 キーワード選択画面

おいたキーワードのみ選択できるようになっている。

3.2.2. 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 3.3 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。

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 に示す。



図 3.3 全体結果表示画面

3.2.3. 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

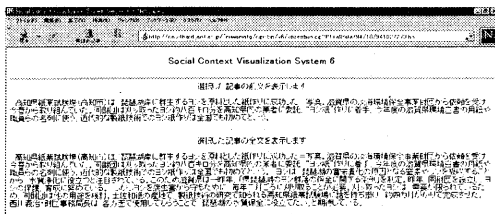


図 3.4 特定記事の内容表示画面

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 [1, 3, 9]。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。

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 3.5 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。

3.3. 再クラスタリング手法

再クラスタリングを実施するには、図 3.6 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 3.7 に示すようになる。ユーザが選択した記事は赤字の selected が付与される。

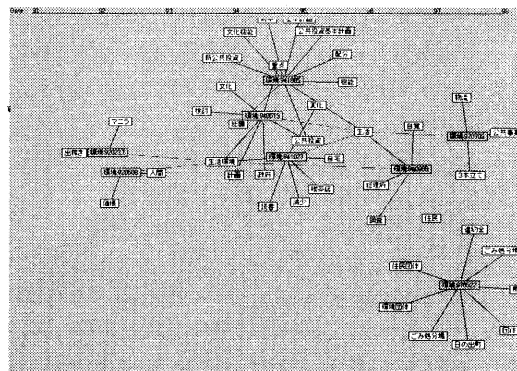


図 3.5 特定クラスタの構造表示画面

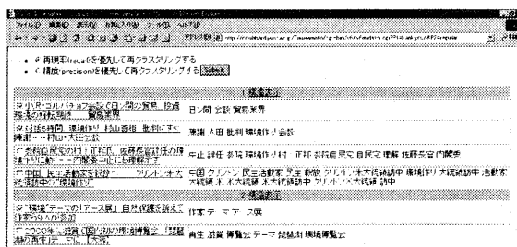


図 3.6 再クラスタリングの選択方法

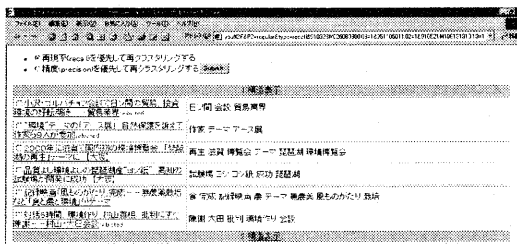


図 3.7 再クラスタリング結果画面

4. 再クラスタリング

4.1. 再クラスタリングの考え方

第 2 節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て再クラスタリングをする必要がある。

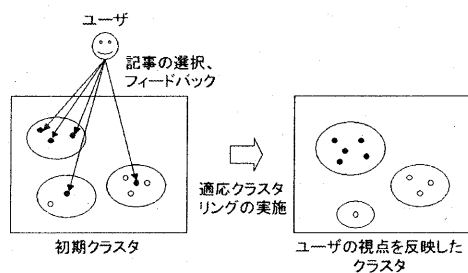


図 4.1 再クラスタリングの概念図

このためには、ユーザからどのようなフィードバック情報を得るか、その結果システムの何を変更するかを決定する必要がある。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。

クラスタを変化させるには索引語の選択とその重みを決定する必要がある。本研究では、索引語の選択に、TFIDF 値[6]、 χ^2 値[5, 8]を用いた方法を利用して、重みを変化させて実験を行った。これらの方法は共起する索引語が前提となるが、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方法についても実験した。図 4.2 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。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した。

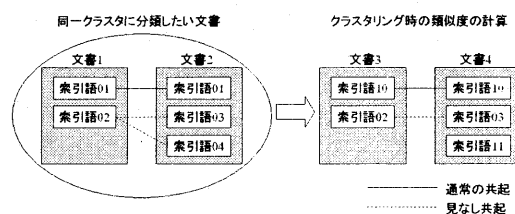


図 4.2 見なし共起を使った類似度の計算

4.2. 再クラスタリング実験

4.2.1. 実験データ

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[7]。この調査をもとに、社会情報の対象として実質的内容

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)	反発反対 (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%)

表 4.1 キーワードと正解セットの特性

*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

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 4.1 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。

4.2.2. 実験手順

実験は以下の手順で行った。

1. 1 つのキーワードを含む記事の集合から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、再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上位から 0% から 100% まで 10% きざみで変化させて実験する。

4.2.3. 実験結果

課題×手法×重みのそれぞれの組み合わせに対して、重要索引語の割合を変化させ、その精度、再現率、 f 値を求めた。図 4.3 は、環境—地球温暖化の課題に対して、見なし共起に基づいて、索引語の割合を変化させたときの精度、再現率、 f 値を示す。このグラフからは、 f 値が 10% 付近で悪化し、その後、50% 付近までは改善するが、それ以上になると、再び悪化することがわかる。10% 付近での f 値の悪化は、再現率が改善しない状態で、精度が悪化したことによるものであり、50% 以降は、再現率の改善が精度の悪化に追いつかないことによるものであることがわかる。

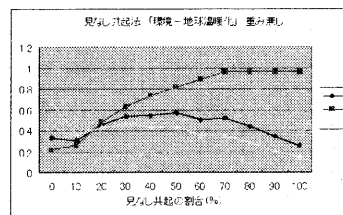


図 4.3 環境—地球温暖化の見なし共起による再クラスタリング結果

表 4.2 には、課題×手法に対して、最大の f 値をとるときの精度、再現率を示した。全般的な傾向として、精度を改善したい場合は、TFIDF 値や χ^2 値を用い、再現率を改善したい場合は、見なし共起を用いるのが良いことがわかる。核—核抑止力、原子力—原発反対

	見なし共起	TFIDF 値	χ^2 値
	精度, 再現率, *1	精度, 再現率, *1	精度, 再現率, *1
核-核抑止力	0.33, 1.0, (1.0, 60, 0.49)	0.22, 0.50, (5.0, 80, 0.30)	0.32, 0.70, (10, 20, 0.44)
原子力-原発反対	0.28, 0.50, (1.0, 20, 0.35)	0.19, 0.42, (1.5, 100, 0.26)	0.19, 0.42, (1.5, 100, 0.26)
環境-地球温暖化	0.44, 0.82, (1.0, 50, 0.57)	0.67, 0.22, (5.0, 0, 0.33)	0.73, 0.30, (5.0, 60, 0.42)
福祉-政策	0.44, 0.80, (5.0, 40, 0.57)	0.57, 0.20, (5.0, 70, 0.30)	0.64, 0.23, (10, 50, 0.33)
通信-無線	0.43, 1.0, (2.0, 60, 0.61)	1.0, 0.23, (5.0, 90, 0.38)	1.0, 0.23, (5.0, 80, 0.38)
情報-情報開示	0.63, 0.74, (5.0, 20, 0.68)	0.87, 0.48, (5.0, 20, 0.62)	0.67, 0.69, (10, 10, 0.68)

表 4.2 課題×索引語決定法に対する精度および再現率

*1: (重み, 重要索引語の割合, f 値)

のように、共起索引語が少なく、また、正解セットを構成する記事に共起索引語が少ない場合には、精度、再現率とも見なし共起による方法が有効であることがわかる。環境-地球温暖化では、TFIDF 値が効いていないが、これは、初期クラスタリングで TFIDF 値を用いて、その際に安定なクラスタができたためではないかと考えている。

5. まとめ

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再クラスタリングについて実験を行った。その実験結果により、共起索引語が少ない場合には、見なし共起を、それ以外の場合、再現率を改善したい場合は、見なし共起を、精度を改善したい場合は、TFIDF 値または χ^2 値を用いるのがよいことが示唆された。

参考文献

- [1] Cutting, D. R. et al. (1992) Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, in Proc. of the 15th Int. ACM/SIGIR Conf.
- [2] 福本, 鈴木 (1999) 語の重み付け学習を用いた文書の自動分類, 情処論, 40(4), 1782-1791.
- [3] Hearst, M. A. et al. (1995) Scatter/Gather as a Tool for the Navigation of

Retrieval Results, In the Working Notes of the 1995 AAAI Fall Symp. on AI Applications in Knowledge Navigation and Retrieval.

[4] CD-毎日新聞 91~97 年版.

[5] 長尾他 (1976) 日本語文献における重要語の自動抽出, 情処論 17(2).

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

[7] 豊浦 (1996) 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.

[8] 渡辺他 (1994) χ^2 法を用いた重要漢字の自動抽出と文献の自動分類, 信学技法 NLC 94-25.

[9] Zamir, O. & O. Etzioni (1999) Grouper: A Dynamic Clustering Interface to Web Search Results, The Eighth Int. World Wide Web Conf.