修 士 論 文

WWWにおける関連リンク集の自動生成

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

田村 雅樹

2006年3月

修 士 論 文

WWWにおける関連リンク集の自動生成

指導教官　白井 清昭 助教授

審査委員主査　白井 清昭 助教授
審査委員　島津 明 教授
審査委員　鳥澤 健太郎 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

410080 田村 雅樹

提出年月: 2006 年 2 月

概 要

ウェブへのアクセスを支援する手法の 1 つにポータルサイトの利用がある．しかし，多種多様なユーザの要求に合ったポータルサイトがウェブ上に存在するとは限らない．したがって，ユーザの興味に応じてポータルサイトを自動的に構築することが望ましい．本研究では，自動的に構築するポータルサイトのコンテンツの 1 つとして関連リンク集を自動的に生成する手法を提案する．その際，特にキーワードの意味の曖昧性と既存のリンク集に対する処理の 2 点に留意する．これに関して，提案手法を実装したシステムを作成し，評価実験を行った．

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

近年のインターネットの普及により，ウェブ上で多種多様な情報を誰もが入手できるようになっている．また，誰もが簡単にウェブサイトを開設できるようになったことで，ウェブ上には膨大な情報が蓄積され，どんどん肥大化している．しかし，ウェブ上の情報は何らかの観点に基づいて整理されているわけではなく，様々なページが無秩序に存在している．ウェブ上には有用な情報がいくつもあるが，様々な情報が蓄積されればされるほど必要な情報を見つけ出すことが困難になる．そこで，多量の情報の中からユーザが必要とするものを容易に探し出すことを支援する技術が望まれる．

現在の情報検索の方法は，Yahoo，Goo，Google といった検索エンジンを用いて求める情報が記述されたウェブページを探すといったものが一般的である．しかし，これら検索エンジンによる検索ではウェブページの内容までは十分に考慮されていない．そのため，ユーザは検索エンジンの検索結果として得られたページを，自分の求める情報が記述されているかどうか 1 つ 1 つ確認していかなければならない．この方法では，たまたますぐに目的とするページが見つかればよいが，そういったことはそれほど多くはなく，ユーザの負担が大きく時間もかかる．

一方，ウェブへのアクセスを支援する手法の 1 つにポータルサイトの利用がある．ポータルサイトとは何かを調べるときに最初に訪れると便利なページであり，Yahoo などの総合ディレクトリサービスやあるテーマに特化した情報を集約したページなどがある．しかし，ユーザの要求は多種多様であり，テーマに合ったポータルサイトがウェブ上に存在するとは限らない．そこで，ユーザが何か調べ物をしているとき，このポータルサイトをユーザの与えたテーマに沿って自動的に構築できれば大変便利である．

1.2　研究の目的

本研究では，ウェブへのアクセスを支援するという観点から，ポータルサイトのコンテンツの 1 つとして関連リンク集の自動生成を目指す．関連リンク集とは，与えたテーマに関する情報が記述されているウェブページを収集し，リンク集として提示するというものである．

関連リンク集を自動生成する際に必要となる処理は大きく分けて以下の 2 点である．

- リンク集に掲載するページの収集・選別
- リンク先ページに関する説明の記述

リンク集を作成するためには前者の処理は必要不可欠である．また，リンク先の内容が分からなければ検索エンジンを利用した場合と同じ問題を抱えることになるため，後者の処理も重要である．本研究では前者に焦点を当てている．

リンク集を作成する際に問題となるのはキーワードの曖昧性である．テーマとして入力されたキーワードに曖昧性があるとユーザの要求を正しく判断できず，テーマに沿っていないページを関連リンク集に加えてしまう恐れがある．そこで本研究ではキーワードの前後の名詞に着目し，同じ意味をもつキーワードを含むページ同士をまとめるクラスタリングを行う．

1.3 本論文の構成

本論文では，2 章ではウェブ上から自動的に情報を取得してして処理を行う関連研究について述べる．3 章では関連リンク集を自動生成する具体的な手法について述べる．4 章では 3 章の手法を実装したシステムの評価実験について述べる．そして 5 章では結論と今後の課題を述べる．

第2章　関連研究

2.1　ウェブページに関する説明の抽出

平野は関連リンク集に掲載するウェブページの説明の情報を他のページから自動的に取得する手法を提案している [1]．この手法では，関連リンク集に提示するページ (対象ページ) の説明を，対象ページへリンクしているページ (参照ページ) の文中から取り出す．

まず，参照ページ中の HTML の要素や対象ページのサイト名を手掛かりとして対象ページに関する記述 (参照箇所) を抽出する．HTML を手掛かりとする手法は，以下の 4 種類のパターンから参照箇所を抽出する．

リスト　アンカーと非アンカーテキストがリスト項目 (`li`) や定義リスト (`dl` 内の `dt` および `dd`) で表されているもの

テーブル　以下のいずれかに当てはまるもの

- アンカーを含むセルの隣のセルに非アンカーテキストが現れるパターン
  (行単位のリンク集)
- アンカーを含む行と非アンカーテキストの行が交互に現れるパターン
  (2 行単位のリンク集)
- アンカーを含むセルが横に並び，その下のセルに非アンカーテキストが現れるパターン
  (列単位のリンク集)

改行　アンカーと非アンカーテキストが改行 (`<br>`) で列挙されているもの

その他　上記に当てはまらないもの

また，サイト名を手掛かりとする手法は，アンカーテキストがサイト名と考えられるならばそれを抽出し，そのサイト名と同じ文字列を含む文を抽出する．

次に，ユーザの利便性を高めるため，抽出した情報をタイプ別に分類・整理する．そこで，抽出した参照箇所にパターンマッチングを行い，以下の 5 種類のカテゴリに分類する．

評価: 利便　ページの利便性，使い勝手，見やすさ等に関する記述

評価: 情報量　ページの規模，情報量等に関する記述

評価: その他　ページの上記 2 つに当てはまらない評価

説明: 機能　ページの機能に関する記述

説明: 記述　ページの一般的な説明

平野の研究では，1.2 節で述べた関連リンク集を自動生成する際に行う処理のうち，後者の「リンク先ページに関する説明の記述」部分に相当する．これは本研究と同様に関連リンク集を生成するための研究であるが，本研究では「リンク集に掲載するページの収集・選別」の処理を行う．したがって，本研究とは相補的な関係にある．

2.2　ウェブディレクトリの自動生成

佐藤らは水族館，動物園，博物館といった特定のカテゴリに関するウェブディレクトリを自動生成する手法を提案している [2]．このシステムは主に Name Collector，Contents Editor，Organizer の 3 つのモジュールからなり，ユーザの入力したカテゴリワードを元にディレクトリを生成する．

Name Collector はカテゴリワードに関するインスタンス (固有名) を収集する．例えば，カテゴリワードとして “aquarium” が与えられた場合，水族館の名前である “Waikiki Aquarium”，“Steinhart Aquarium” や “Monterey Bay Aquarium” 等を収集する．インスタンスとして収集するものは以下の 2 つである．

- カテゴリワードを主要辞とする固有名
- インスタンスが現れたリスト内のその他の固有名

実際には，インスタンスはリンク集から抽出した固有名と URL の組からなる．

Contents Editor は各インスタンスのダイジェストページを作成する．まず，集めたインスタンスの URL から名前，住所，市町村コード，電話番号，概要といった要約情報の抽出を行う．この要約情報を元にインスタンスのダイジェストページを生成する．ダイジェストページには名前，住所，電話番号，概要，リンク集が記載される．

Organizer は 2 つのタスクを実行する．1 つ目は各インスタンスの同一性チェックである．2 つのインスタンスについてインスタンス固有の情報を比較し，同一と判断された場合はインスタンスをマージする．インスタンス固有の情報とは，市町村コード，名前，住所，電話番号の 4 つである．例えば，名前が異なっていても住所が一致すれば同一と判断する．2 つ目のタスクは地方別に階層構造で整理した目次ページを生成することである．地方は市町村コードから求め，階層の深さはインスタンス数から定まる．

佐藤らの研究では本研究と同様にユーザの入力したキーワードを元にしてページを自動生成する．しかし，生成対象がウェブディレクトリであり，本研究のように関連リンク集を生成するわけではない．

2.3 ウェブページのクラスタリング

Zamir らは文書間で共通するフレーズに基づいた Suffix Tree Clustering(STC) アルゴリズムによるクラスタリング手法を提案している [3]．Suffix Tree とは，文字列 (以下 S) から全ての接尾辞を取り出してコンパクトなトライ構造としてまとめたものであり，以下のような性質をもつ．

1. ルートをもつ有向グラフ
2. 内部ノードは 2 つ以上の子をもつ
3. 各枝はラベルとして S の空でない部分文字列をもつ (トライ構造)
4. あるノードに接続しているノードは全て異なるラベルをもつ (コンパクト)
5. S の任意の接尾辞 s について，s と同一のラベルをもつ接尾辞ノードが存在する

作成した Suffix Tree のノードはベースクラスタを表す．

STC では，まず各文書の一部から Suffix Tree を構築し，ベースクラスタを作成する．その後，文書集合の重複が多いクラスタ同士をマージしてクラスタを大きくする．でき上がったクラスタのうち，クラスタ内の文書数やフレーズ内の単語数から計算されるスコアの上位 10 件を最終的に取り出す．

STC によるクラスタリングでは検索エンジンの検索結果をユーザに理解しやすい形にして出力する．ウェブページのクラスタリングによってユーザの利便性を高めるという点では本研究と同様だが，STC が従来のトピックのクラスタリングを行う点に対し，本研究ではキーワードの曖昧性に着目したクラスタリングを行うことで関連リンク集を作成する．

また，金子らは検索キーワードを問わず頻出する「基本カテゴリ」のうち，指定したもののみを採用する検索エンジン “Gooots” を提案している [4]．この基本カテゴリは以下の 5 種類である．

- 特定のアプリケーションを必要とするページ (pdf，doc など)
- ショッピングサイト
- 書籍
- 掲示板，日記
- シラバス

また，検索目的として以下の 2 つに注目している．

- 仕事や学習のための検索

- 商品購入のための検索

カテゴリの選択では，Gooots 通常検索と Gooots 目的別検索という，選択方式の異なる 2 種類のシステムを作成している．Gooots 通常検索ではユーザが基本カテゴリから採用するカテゴリを手動で選ぶ．そして，選択されたカテゴリに属するページのみをまとめ，どのカテゴリにも登録されなかったページはリスト表示する．Gooots 目的別検索では検索目的を選択するとその検索目的に合わせて自動的にカテゴリを選択する．例えば，基本カテゴリ「特定のアプリケーションを必要とするページ」は学習目的で検索されることが多く買い物目的であることはほとんどない，といった情報を利用している．

クラスタリングは，まず googleAPI を用いて検索を行いウェブページを取得する．そして，経験則によって決定された特徴語を含むページを選択されたカテゴリに登録することで実現する．

金子らの研究ではクラスタリングにおけるカテゴリが基本カテゴリ 5 種類に固定されているが，本研究では検索キーワードに応じて動的にクラスタリングを行う．

Vivísimo[5] は検索結果を動的にクラスタリングしてカテゴリ分けを行うメタサーチエンジンの Clusty.com (日本語版 Clusty.jp，2006 年 2 月現在 β 版) を提供している．詳細は不明だが，Clusty は “Velocity” と呼ばれる独自クラスタリングエンジンを用いてカテゴリ分類を行う．例えば,「英会話」をキーワードとして与えると表 2.1 のようにカテゴリ分類される (2006/02/03 現在).

表 2.1: Clusty によるクラスタリング例

クラスタ	件数
(すべての結果)	154
英会話学校	26
英語・英会話	26
子供	22
初心者英会話	16
英会話スクール・英会話教室	11
⋮	⋮

Clusty は本研究のクラスタリングと近い立場にあるといえる．しかし，本研究のようにキーワードの前後の文脈が異なる場合でも同じ意味をもつキーワードを含むページを 1 つのクラスタにまとめることまでは試みていないようである．

第3章　提案システム

3.1　システムの概要

本研究で構築するシステムは，ユーザによって与えられたテーマに関連するページをウェブ上から検索し，リンク集を自動的に作成する．また，キーワードの前後の名詞に着目したクラスタリングを行い，テーマの曖昧性を考慮した適切なリンク集を作成する．

3.2　処理の流れ

本システムで行う処理は全部で 6 段階の部分処理に分かれる．それぞれの処理の内容は以下の通りである．

1. テーマの入力
2. 候補ページの取得
3. 候補ページの追加
4. 不要なページの削除
5. クラスタリング
6. リンク集の出力

3.2.1　テーマの入力

まず，リンク集のテーマを決める．具体的には，ユーザに作成するリンク集のテーマをキーワードとして入力させる．テーマは単一あるいは複数の名詞とする．

3.2.2　候補ページの取得

膨大な数のウェブページ全てについて処理を行うのは不可能であるため，まず最低限の選別を行う．図 3.1 に候補ページを取得する手順を示す．3.2.1 項で入力されたテーマを検索キーワードとして検索エンジン Goo でウェブ検索を行い，上位 500 件以内に現れたペー

ジをリンク集に掲載するページの候補として取得する．ページの取得にはGNU Wgetを用いる．以下，リンク集に掲載するページの候補を候補ページと呼ぶ．特にこのステップで取得したページを初期候補ページと呼ぶ．

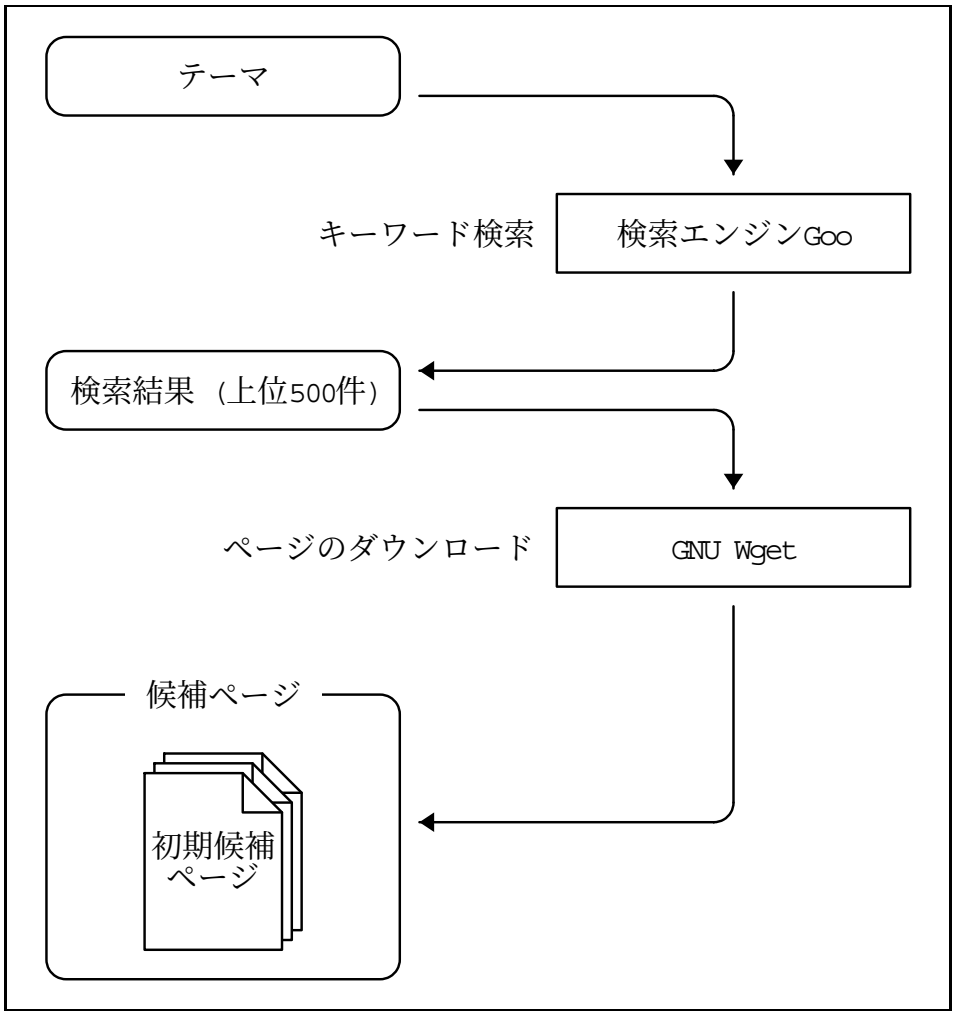


図 3.1: 候補ページの取得

Gooの検索結果はキャッシュとして保存されているため，実際にページを取得するときにはそのページが消えていたりサーバが停止していることがある．また，Gooの検索結果にはテキストファイルやPDFファイルなどのHTMLでないファイルも含まれている．そこで，ページを取得する前にサーバレスポンスを調べる．エラーが発生したりContent-Typeが “text/html” か “application/xhtml+xml” でないものはダウンロードしない．

3.2.3 候補ページの追加

3.2.2 項で得た候補ページにリンク集が含まれていた場合，そのリンク集には，Goo の検索結果の上位 500 件以内に存在しないがリンク集に加えるべきページへのリンクが含まれていることがある．すなわち，そのリンク集は作成する関連リンク集のサブセットとなっていることがある．そこで，図 3.2 に示すように候補ページのリンク集からリンク先のページを取得する．

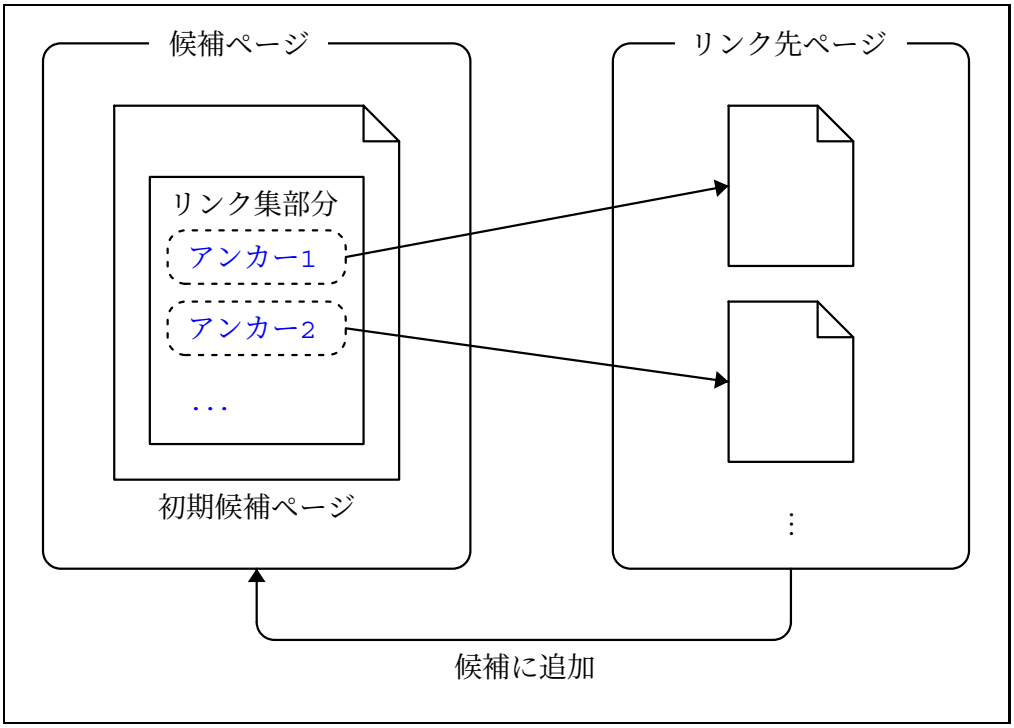


図 3.2: 候補ページの追加

リンク先に加えるページはテーマと関連していなければならない．そこで，以下の手順でリンク集内のテーマと関連があると考えられるページのみを候補ページに追加する．

1. 各候補ページ内のリンク集部分を検出・取得する．具体的なリンク集検出アルゴリズムについては 3.3 節で述べる．

2. 取得したリンク集のうち，テーマと明らかに関連がないものを除く．ここでは，キーワードが 1 つでも含まれていれば関連があると考え，全てのキーワードが含まれていないリンク集のみを除く．

3. 残ったリンク集のリンク先ページを取得する．

4. 取得したページのうち，全てのキーワードを含むものを候補ページに追加する．以下，ここで追加したページを追加候補ページと呼ぶ．

上記の手順に従い，ページ全体がリンク集となっているものだけではなく，一部のみがリンク集となっているページに関してもリンク集部分を抽出する．これは，ページのコンテンツに関連するページ，ページ作成の際に参考にしたページ等のリンク集が存在するためである．WikiPedia の Perl の項の例を図 3.3 に示す．図の枠線部分 (「外部リンク」の下) は Perl と関連するページへのリンク集である．

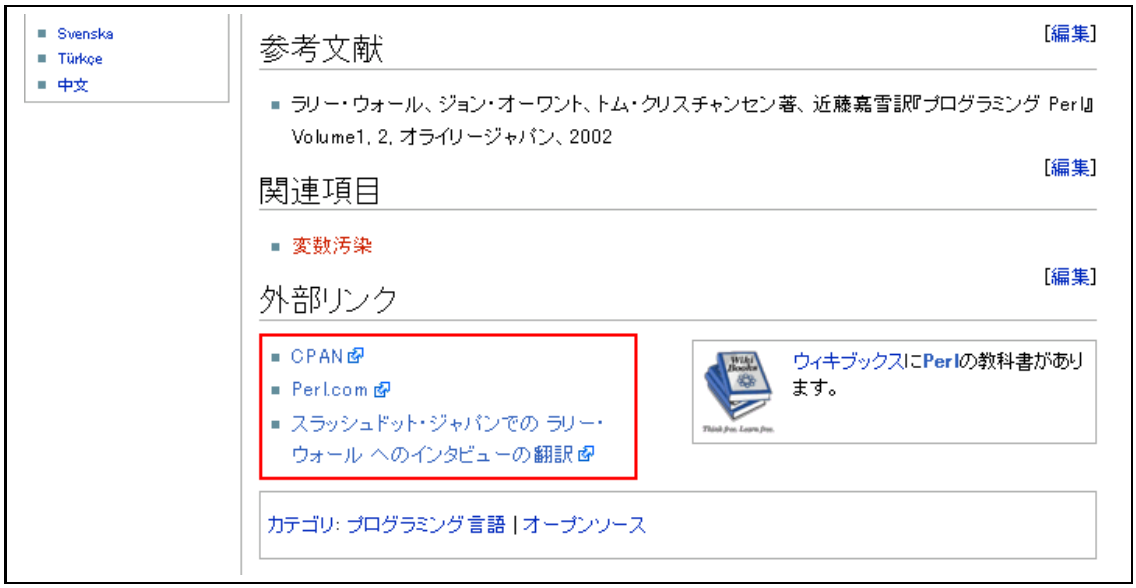


図 3.3: 関連するページへのリンク集の例
(WikiPedia > Perl の項 より)

3.2.4 不要なページの削除

関連リンク集から別のリンク集に飛び，そこから更にリンクを辿るのは二度手間となり，ユーザにとって不便である．そこで，以下の手順でリンク集のみからなると考えられるページを候補ページから除外する．

1. 各候補ページ内のリンク集部分を検出・取得する (3.3 節参照).
2. 各候補ページ全体とリンク集部分の HTML データをそれぞれプレインテキストに変換する．
3. リンク集部分のプレインテキストがページ全体のプレインテキストの 80%以上を占めていたらリンク集のみからなるページと判断し，そのページを候補ページから除外する．

プレインテキストで比較するのは，製作者によって HTML のマークアップが異なるため，タグの数などで結果が変わるのを防ぐためである．

このアルゴリズムではページの一部のみがリンク集となっているようなものは除外されない．これは，リンク集部分でないページの主要な部分が関連リンク集に掲載すべき内容である可能性があり，そういったページまで除外してしまうのを防ぐためである．

3.2.5 クラスタリング

ユーザの入力するテーマは名詞の集合のみであるため，テーマに曖昧性があるとユーザの意図を正しく判断するのは難しい．例えば，図 3.4 に示したように，テーマが「松井」の場合は「松井秀喜」や「松井証券」など複数の「松井」が考えられ，ユーザの意図を判断できない．そこで，キーワードの曖昧性を考慮したクラスタリングを行い，各クラスタごとにリンク集を作成する．テーマが「松井」の場合は図 3.5 のように松井秀喜に関するリンク集や松井稼頭央に関するリンク集などが作成される．作成したクラスタ (リンク集) にはキーワードの曖昧性に着目して適切な名前を付ける．

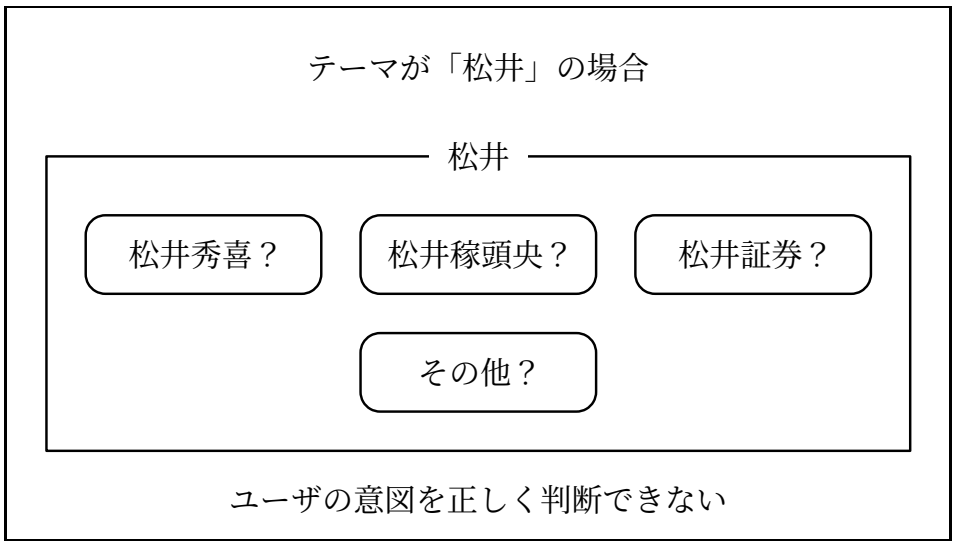


図 3.4: テーマの曖昧性

従来のクラスタリングでは同じトピックのページ同士を同じクラスタとしてまとめるが，本研究ではキーワード前後の名詞に着目し，同じ意味を持つキーワードを含むページをリンク集としてまとめる．具体的なアルゴリズムは 3.4 節で述べる．

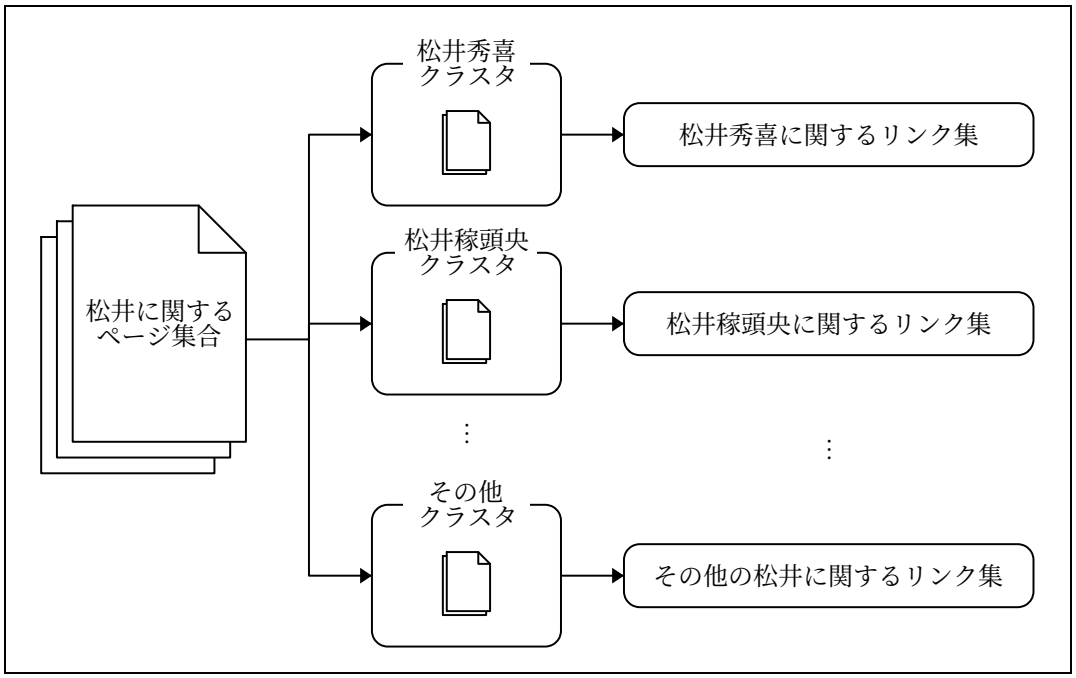


図 3.5: クラスタリング

3.2.6 リンク集の出力

クラスタリングの結果得られたリンク集を出力する．ユーザが自分の要求に適したクラスタを選択することで，そのクラスタに属するページへのリンク集を表示する．なお，出力は現時点では未実装である．

3.3 リンク集の検出

本節ではウェブページ内のリンク集部分を検出する手法を紹介する．提案手法ではHTMLの構造に注目しており，リンクが特定の形式で列挙されている場合にリンク集だと判定する．処理の流れを以下に述べる．

1. リンクがリスト形式で列挙されている部分を抽出する．
2. リンクが改行によって列挙されている部分を抽出する．
3. リンクが表形式で列挙されている部分を抽出する．
4. 1〜3で抽出したリンク集から重複しているものを除く．

3.3.1 リスト形式のリンク集の検出

リスト形式のリンク集とは，リンクがリストの項目 (li 要素) の内容として列挙されているものを指す．図 3.6 に例を示す．

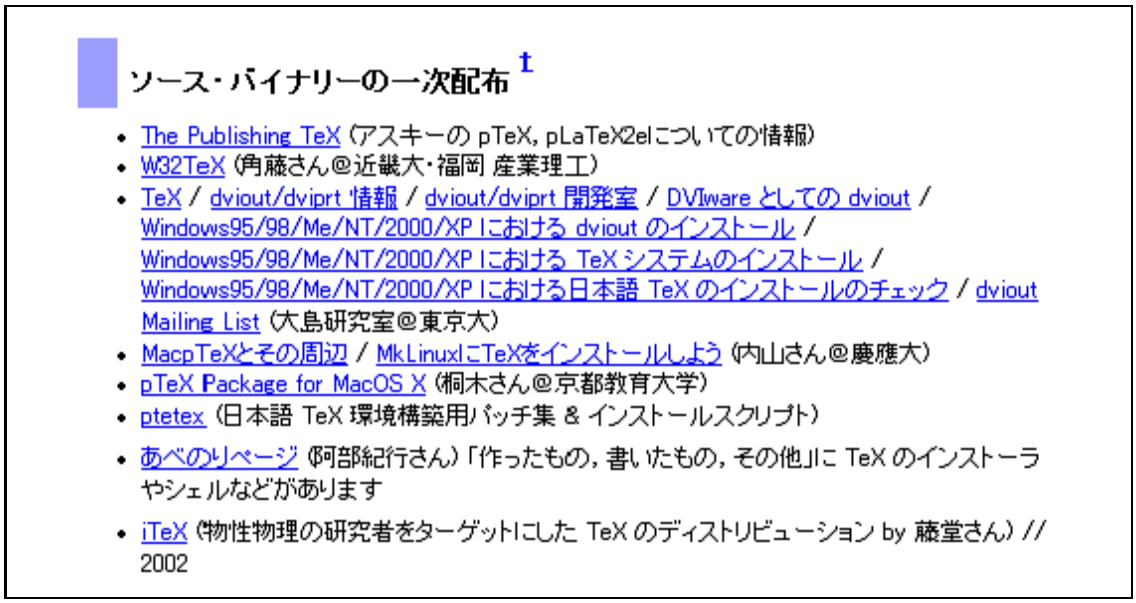


図 3.6: リスト形式のリンク集
(TEX Wiki > 国内リンク より)

リストによってリンク集が表されているページの多くには以下のような特徴がある．

- 順序なしリスト (ul) であり，順序付きリスト (ol) は少ない．
- タグの直後にアンカーが現れる．
- リンク先ページの説明の有無，及び説明の書き方はページによる．

また，リンク集であるからリスト内には外部へのリンクが一定数必要であると考えられる．そこで，リスト型のリンク集は以下のように判断する．

> タグとそれに対応する
タグ間で，
>
> - タグの直後に<a>タグが存在する．
> - その<a>タグの href 属性が外部リンクになっている．
>
> を満たすものが閾値以上存在すればリスト型のリンク集である．

いくつかのページを調べたところアンカーが 2 個以下でリンク集というものが存在した．しかし，そういったリンク集はそれほど多くはないことと，リンク集ではない本文中のリストもリンク集として判断してしまう恐れがあるため，閾値は 3 とする．

その他，定義リスト (dl) でリンク集が表されているページも存在する．しかし，本研究では定義リストの判定方法は実装していない．

外部リンクの判定

外部リンクの判定は以下のようにして行う．

(a) 相対パスで表される URL は内部リンク

(b) そのページの URL とリンク先ページの URL を比較して，ドメインとその直下のディレクトリ名が同じならば内部リンク

(c) その他の URL は外部リンク

ここで，(a) の相対パスで表される URL とは，“./”，“../”，“/” で始まる URL を指す．ただし，単に相対パスを除くだけではメール送信アンカー (“mailto:” で始まるもの) 等が含まれてしまう．そういったものを除外するため，実装では “http://” または “ftp://” で始まるもののみを絶対パスとして考える．また，(b) のドメインとその直下のディレクトリとは，例えばページの URL が “http://www.jaist.ac.jp/is/index-jp.html” ならば “http://www.jaist.ac.jp/is/” の部分を指す．ここで，ドメインだけでなく直下のディレクトリも含めて判断するのは，ドメインだけでは外部サイトを判断できないケースがあるためである．例えば，Yahoo!ジオシティーズのサービスを用いているサイトで考えると，“http://www.geocities.jp/abc/” から “http://www.geocities.jp/def/” へのリンクは外部サイトへのリンクとするべきである．しかし，ドメインだけで判断した場合は内部サイトへのリンクと扱われてしまう．

3.3.2 改行によるリンク集の検出

改行によるリンク集は，アンカーの直後に改行 (br 要素) が現れ，それが繰り返されているものを指す．図 3.7 に例を示す．

改行によるリンク集は以下のように判断する．

図 3.7: 改行によるリンク集
(フィギュアスケート通信 より)

href 属性が外部リンクである<a>タグに対して，

- 対応するタグの直後が
  1. 0 個以上のインライン要素のタグ
  2.
タグ

  の順番で並んでいる．
- 上記の
タグの直後が
  1. 0 個以上のインライン要素のタグ
  2. 次の<a>タグ

  の順番で並んでいる．

を満たし，それが閾値以上の回数で連続して出現すれば改行によるリンク集である．

閾値はリスト形式と同様に 3 とする．また，インライン要素とはここでは以下のものを指す．

- 文字レイアウト指定要素 (font，b，i 等)
- テキスト修飾要素 (em，strong，q 等)
- 画像 (img)

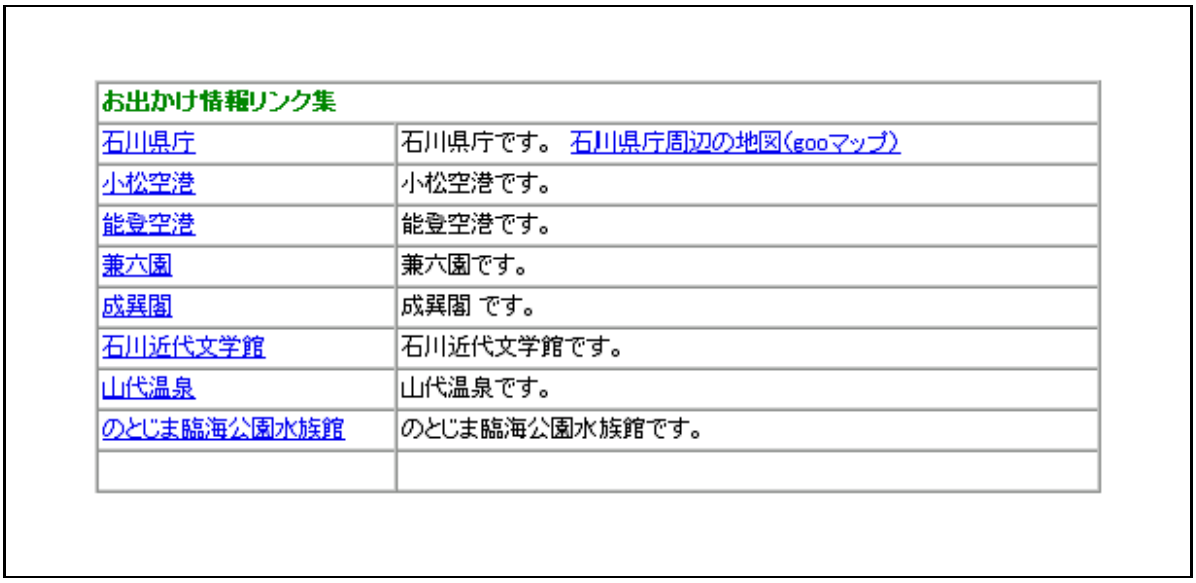
インライン要素のタグを含んでもよいことになっているが，これは以下のようなリンクもリンク集を構成するリンクの 1 つと判断するためである．

```
<b><a href="url">リンク先ページ名</a></b><br>
```

3.3.3 表形式のリンク集の検出

表形式のリンク集は表内に行単位でリンクが列挙されているものである．行単位としたのは，各行がリンクのみの 1 列のもの，またはリンクとリンク先ページの説明がある 2 列のものが多いためである．図 3.8 に例を示す．

行単位で考えるため，表形式のリンク集は以下のように判断する．



お出かけ情報リンク集	
石川県庁	石川県庁です。 石川県庁周辺の地図(gooマップ)
小松空港	小松空港です。
能登空港	能登空港です。
兼六園	兼六園です。
成巽閣	成巽閣 です。
石川近代文学館	石川近代文学館です。
山代温泉	山代温泉です。
のとじま臨海公園水族館	のとじま臨海公園水族館です。

図 3.8: 表形式のリンク集
(shirabeyou.com > 石川県の地図・交通・観光リンク集 より)

<table>タグとそれに対応する</table>タグ間で,

- <tr>タグとそれに対応する</tr>タグ間で,
  - <a>タグが存在する.
  - その<a>タグの href 属性が外部リンクになっている.

  を満たすものが存在する.
- その<tr>タグが閾値以上存在し，表全体の行数の 8 割以上を占める.

を満たせば表形式のリンク集である.

閾値は上記 2 つのリンク集と同様に 3 とする.

3.3.4 重複したリンク集の削除

上記 3 種類のリンク集はそれぞれ別々に検出するため，重複して検出される場合がある．そこで，あるリンク集が別のリンク集の一部となっていた場合，一部となっていた方のリンク集を削除する.

図 3.9 は表型のリンク集の中にリスト型のリンク集が含まれている例である．この場合

は1行目のリスト型のリンク集部分が表型のリンク集の一部となっているため，リスト型のリンク集を削除して表型のものだけを残す．

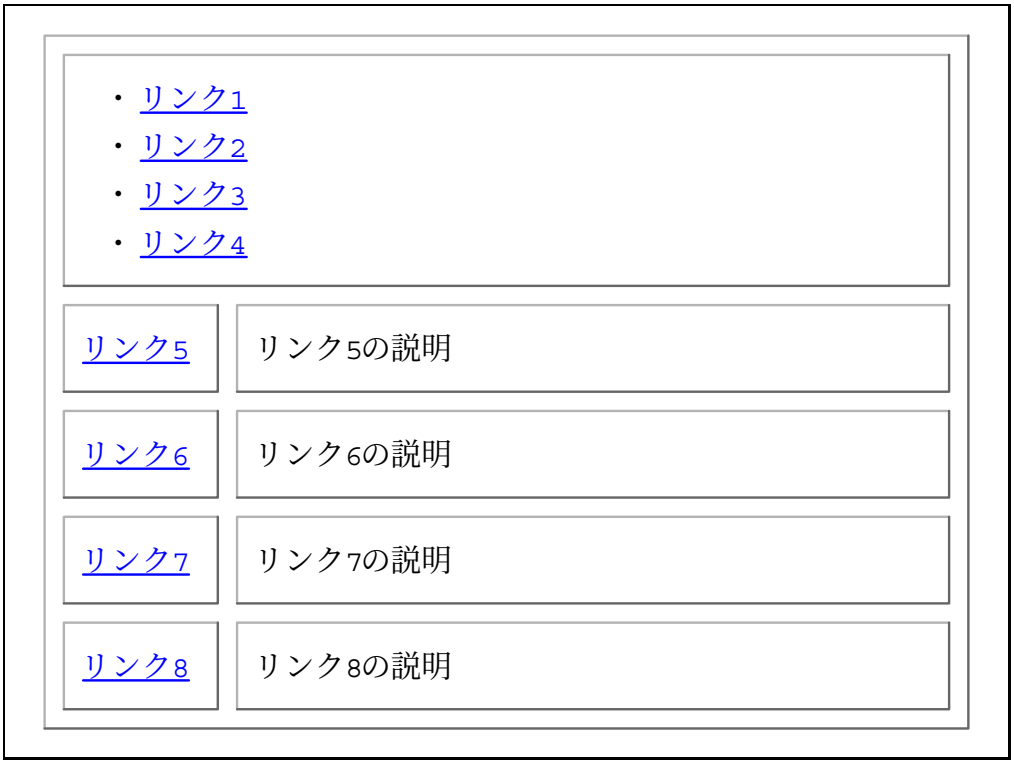


図 3.9: リンク集が重複している例

3.4 キーワードの曖昧性を考慮したクラスタリング

本節ではキーワードの曖昧性を考慮したクラスタリングのアルゴリズムについて述べる．

本研究で行うクラスタリングはキーワードの曖昧性を考慮するという特殊なものである．そのため本研究では，最初にいくつかのページからベースとなるクラスタを作成し，その後で残ったページをクラスタに追加していくという，トピックごとにまとめる従来のクラスタリングとは異なる手法をとる．なお，キーワードが複数個与えられた場合はそれぞれのキーワードごとに独立してクラスタを作成する．

3.4.1 基本クラスタの作成

まず，ベースとなるクラスタ (基本クラスタ) を作成する．このクラスタは，キーワードの意味の異なるページが別々のクラスタになるように作らなければならない．

例えば，キーワードが「松井」の場合は「松井 秀喜」や「松井 稼頭央」に関するページをまとめたものなどがクラスタの候補として挙げられる．また,「野球」の場合は「高校 野球」や「プロ 野球」に関するページをまとめたものなどが候補になる．このように，キーワード前後の単語，特に名詞が異なれば意味も異なると考え，前後に同じ名詞が現れるキーワードを含むページを 1 つのクラスタとすることは有望である．そこで，本手法では以下の手順で基本クラスタを定める．

1. 形態素解析を行い，キーワードの直前・直後の名詞を抽出する．
2. 各ページごとにキーワードの意味を同定する．
3. キーワードの意味が同じページの個数を数え，名詞表を作成する．
4. ページ数の多いものを基本クラスタとする．

名詞の抽出

まず，形態素解析ツールである茶筌 [6] を用いて形態素解析を行う．茶筌の結果は，デフォルトでは数字や英字は 1 文字で 1 つの単語 (形態素) とみなされる．しかし，連続する英字や数字は 1 つの単語であることが多い．そこで，それらの語を連結して英字列や数字列，英数字列とする．また，キーワードに複合名詞が含まれる場合も結果を連結し，その複合名詞を 1 つの単語とみなす．

形態素解析を行った後，キーワードの前後の名詞，すなわちキーワード直前・直後の名詞の組を抽出する．キーワードが複数与えられている場合は各キーワードごとに別々に名詞組を取り出す．また，同じキーワードが複数回出現している場合はそれら全てについて名詞組を取り出す．キーワードの直前または直後の単語が名詞でない場合，あるいは単語が存在しない場合は「[名詞なし]」とする．図 3.10 に名詞の抽出例を示す．キーワードが「松井」で「ヤンキースの松井秀喜選手」という部分が現れたときの「松井」の前後の単語は「の」と「秀喜」である.「の」は助詞,「秀喜」は名詞であるため，抽出する名詞組は ([名詞なし], 秀喜) となる．

英字列などについては茶筌では品詞を判別することができないが，日本語の名詞に隣接する英字列や英数字列は名詞として使われている可能性が高いと考えられる．したがって，キーワードに隣接する英字列・英数字列は名詞として扱い，抽出の対象とする．ただし，数字のみからなる数列は名詞としてふさわしくないため，抽出の対象とはしない．また，英語のストップワードリストを用いて，明らかに名詞ではないと断定できる語や役に立たないであろう語は抽出の対象としない．このストップワードリストには [7] を用いた．

キーワードの意味の同定

ページごとに，そのページ内で使われているキーワードがどの意味をもつのかを決める．キーワードの意味とは，例えばキーワードが「松井」ならばそれが「松井秀喜」な

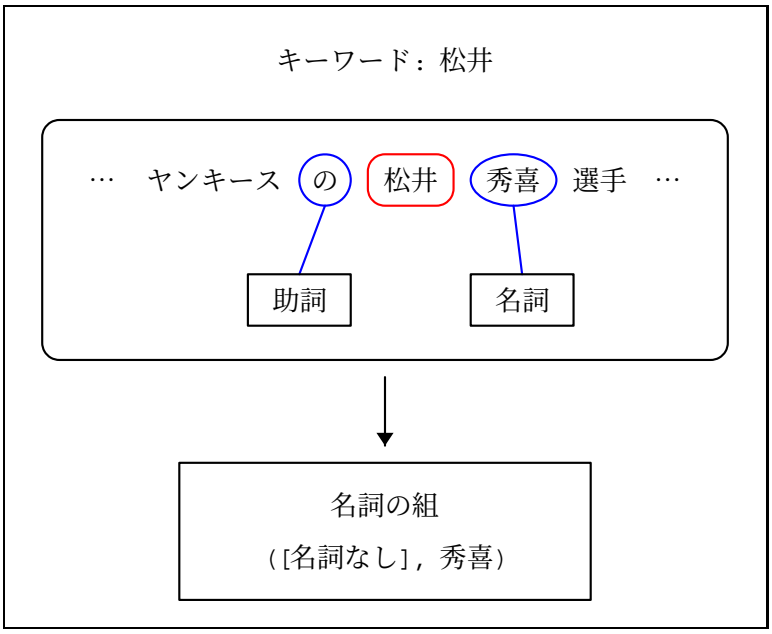


図 3.10: 名詞の抽出例

のか「松井稼頭央」なのかといったように，その「松井」が具体的に表しているものを指す．ここでは，抽出されたキーワードと前後の文字列でキーワードの 1 つの意味を表すとみなす．図 3.10 の例で考えると，キーワード「松井」に対して抽出される名詞組は $([\text{名詞なし}], \text{秀喜})$ であるから，そのキーワードの意味は「松井秀喜」であるとみなす．

ページ p におけるキーワード k の意味を $mean(p, k)$ とする．p に k が 1 回だけ現れ，その前後の名詞組が (n_p, n_f) である場合，$mean(p, k) = (n_p, n_f)$ とする．p に k が複数回現れるが，前後の名詞組が全ての k について (n_p, n_f) の 1 組だけの場合も $mean(p, k) = (n_p, n_f)$ とする．

p に k が複数回表れ，前後の名詞組が 2 組以上ある場合は，もっとも出現頻度の高い名詞組を $mean(p, k)$ とする．しかし，ページによっては複数のテーマについて記述されていることがあり，そういったページに対してはもっとも出現頻度の高い名詞組のみに意味を限定するのは好ましくない．そこで，出現回数が最大値の 75%以内である名詞組もキーワードの意味として選ぶことにする．同じページに対して，あるキーワードの意味が複数個選ばれた場合，そのページは複数の基本クラスタに属する可能性がある．

あるページ p_x に対して，キーワード「松井」の前後の名詞の出現回数が表 3.1 であったとする．この場合，まずもっとも出現頻度の高い $([\text{名詞なし}], \text{秀喜})$ がキーワードの意味として選ばれる．また，$([\text{名詞なし}], \text{秀喜})$ の出現回数は 4 回であるため，その 75%(3 回) 以上の出現回数である $([\text{名詞なし}], \text{稼頭央})$ もキーワードの意味として選ばれる．したがって，$mean(p_x, \text{松井}) = (([\text{名詞なし}], \text{秀喜}), ([\text{名詞なし}], \text{稼頭央}))$ となる．

表 3.1: キーワード前後の名詞の出現回数

直前の名詞	直後の名詞	出現回数
[名詞なし]	秀喜	4
ヤンキース	秀喜	1
[名詞なし]	稼頭央	3

名詞表の作成

キーワードの意味を同定することで，各ページについて前後の名詞の組が定まる．その結果，名詞組について各ページをまとめることで名詞組とページの集合に対応した表が作成できる．このとき，表をページ数 (ページの集合の要素数) について降順にソートしておく．キーワードが「松井」の場合の例を，ページ数の上位 6 件まで表 3.2 に示す．なお，ページ集合全ては書けないため，Goo の検索結果で上位だった 2 ページ分 (番号表記) とページ数を示した．

表 3.2: 名詞表の例

名詞組	ページ集合	ページ数
([名詞なし], [名詞なし])	(003, 004, ...)	211
([名詞なし], 秀喜)	(002, 016, ...)	102
([名詞なし], 稼頭央)	(006, 018, ...)	23
([名詞なし], 雄飛)	(240, 312, ...)	22
([名詞なし], 秀)	(043, 052, ...)	15
([名詞なし], 大輔)	(005, 008, ...)	7
⋮	⋮	⋮

基本クラスタの決定

名詞表からクラスタとするのに適した名詞組を選び，基本クラスタとする．

キーワードの意味として選ばれたある名詞組について，その名詞組が選ばれたページの集合の数が少ない場合，それらのページ集合は例外的なものについて述べたものであると考えられる．したがって，そういったページの集合は基本クラスタよりもその他のページ群とする方が好ましいといえる．また，名詞組が ([名詞なし], [名詞なし]) となっているものは前後の名詞からではキーワードの意味を判断できなかったものであるため，基本クラスタとするには適さない．そこで，基本クラスタを以下のように決定する．

1. 名詞表からページ数が上位 5 件以内でないものを除外する．また，ページ数が最上

位のページ数の 10%未満であるものを除外する．

→ 例外の削除

2. 名詞組が ([名詞なし], [名詞なし]) であるものを除外する．

→ 意味を判断できなかったものの削除

3. 残ったものを基本クラスタとする．以下，この時点で基本クラスタに属しているページを基本ページと呼ぶ．

ここで，キーワードの意味から基本クラスタの名前を決める．この名前はキーワード直前の名詞，キーワード，キーワード直後の名詞を連結したものとする．ただし，キーワード直前または直後の名詞が「[名詞なし]」の場合はそれを連結の対象とはしない．例えばキーワードが「松井」でキーワードの意味が ([名詞なし], 秀喜) の場合は「松井秀喜」クラスタと名付ける．

最後に，基本クラスタに属さなかったページの集合をまとめて「その他」クラスタとする．

例

表 3.2 の名詞表を元にクラスタを作成することを考える．

まず，ステップ 1 によって下位の名詞組を除外する．ページ数が上位 5 件以内であるのは「([名詞なし], 秀)」までである．ページ数の最上位は「([名詞なし], [名詞なし])」であり，その数は 211 であるから，10%以内であるのは「([名詞なし], 雄飛)」までである．したがって，両方を満たすのは以下の 4 組である．

- ([名詞なし], [名詞なし])
- ([名詞なし], 秀喜)
- ([名詞なし], 稼頭央)
- ([名詞なし], 雄飛)

次に，ステップ 2 によって名詞組「([名詞なし], [名詞なし])」を除外する．その結果，以下の 3 組が残る．

- ([名詞なし], 秀喜)
- ([名詞なし], 稼頭央)
- ([名詞なし], 雄飛)

したがって，最終的に基本クラスタは以下の 3 つになる．

- 「松井秀喜」クラスタ
- 「松井稼頭央」クラスタ
- 「松井雄飛」クラスタ

3.4.2 基本クラスタへのページの追加

前項で基本クラスタに属さなかったページはキーワードの意味が不明だったものとして扱われる．しかし，例えば以下のような理由から，本来は基本クラスタに属すべきページがその他のページ集合として扱われてしまった可能性がある．

- キーワード前後の名詞が (文脈から人間には分かる程度に) 省略されていた

  例: 本文が「ヤンキースの松井は...」
  「ヤンキース」から松井秀喜だと分かるが,「松井」の前後の単語は「の」と「は」であるしたがって，名詞組は ([名詞なし], [名詞なし]) となり，([名詞なし], 秀喜) とは別のクラスタに属することになる．

- 前後の名詞に誤字があった

  例: 本文が「ヤンキースの松井秀樹選手は...」
  「松井」の前後の単語は「の」と「秀樹」である．したがって，名詞組は ([名詞なし], 秀樹) となり，([名詞なし], 秀喜) とは別のクラスタに属することになる．

- 前後の名詞にノイズとなるような語が混じっていた

  例: 本文が「ヤンキース松井秀喜...」
  「松井」の前後の単語は「ヤンキース」と「秀喜」である．したがって，名詞組が (ヤンキース, 秀喜) となり，([名詞なし], 秀喜) とは別のクラスタに属することになる．

そこで，基本クラスタと基本クラスタに属していないページ間の類似度を計算し，類似度の高いページを基本クラスタに追加する．

ページ追加の手順は以下の通りである．

1. 各「その他」クラスタに含まれるページおよび「その他」クラスタを除いた各基本クラスタの単語ベクトルを定める．単語ベクトルの作成方法は後述する．
2. 「その他」クラスタに含まれるページと「その他」クラスタを除いた基本クラスタ全ての組み合わせについて類似度を計算する．
3. 類似度の最も高いページと基本クラスタの組について，

(a) 類似度が閾値以上ならばそのページをそのクラスタに追加する．また，そのページを「その他」クラスタから除外する．

(b) 類似度が閾値未満ならば処理を終了する．

4. 「その他」クラスタにページが存在していれば 2〜3 の処理を繰り返す．

ステップ 2 の類似度計算にはコサイン類似度を用いる．すなわち，単語ベクトル $\boldsymbol{a}$ と $\boldsymbol{b}$ の間の類似度は以下の式で与えられる．

$$s(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}^T \cdot \boldsymbol{b}}{||\boldsymbol{a}|| \times ||\boldsymbol{b}||} \tag{3.1}$$

$||\boldsymbol{x}||$: 単語ベクトル$\boldsymbol{x}$のユークリッドノルム

ウェブ文書のクラスタリングでは，コサイン類似度はユークリッド距離を用いた方法などと比べて性能がよい [8]．また，ステップ 3 の類似度の閾値は 0.4 とする．この値を用いるのは予備実験で比較的良い精度が得られたためである．

単語ベクトル

キーワードから離れた場所に出現する単語がそのキーワードの意味を表していることはまずないと考えられる．そこで，キーワードおよびその周り 50 単語以内に出現する自立語をページの単語ベクトルの要素 (素性) とする．3.4.1 項の名詞の抽出処理のときと同様に，ストップワードを除いた英字列や英数字列は名詞とみなす．なお，単語の出現位置は考慮しない (Bag of words)．

一般に，情報検索の分野では不必要な単語と重要な単語に差異を持たせるため，単語に重み付けを行う．この重み付けには TF(Term Frequency) 値と IDF(Inverse Document Frequency) 値の積である TF-IDF 値がよく用いられている．IDF 値は各文書を特徴付ける重みであり，多くの文書で頻出する語の重みを小さくし，少数の文書にしか出現しないような語の重みを大きくする．しかし，本研究においては各文書を特徴付ける語よりも各クラスタを特徴付ける語に大きな重みを与える方が望ましい．そこで，クラスタを特徴付ける重みとして ICF(Inverse Cluster Frequency) 値を導入する．

単語 t の ICF 値は以下の式で定義される．

$$icf(t) = \ln\left(\frac{N_c}{cf(t)} + 1\right) \tag{3.2}$$

N_c : クラスタの総数

$cf(t)$: 単語 t が出現するクラスタの個数

(N_c, $cf(t)$ は「その他」クラスタも考慮に入れる)

ICF 値は IDF 値を文書単位からクラスタ単位に直した値であるといえる．

したがって，ページ p における単語 t の重み $w(t,p)$ は以下の式で与えられる．

$$\begin{aligned} w(t,p) &= Normalize\big(tf(t,p)\big) \times icf(t) \\ &= \frac{tf(t,p)}{\sum_t tf(t,p)} \times \ln\left(\frac{N_c}{cf(t)} + 1\right) \end{aligned} \tag{3.3}$$

式 (3.3) では，TF 値をページ p の素性の数で割ることで正規化を行っている．これは，単語の出現回数をそのまま用いると ICF 値と比べて重みが大きくなりすぎるためである．

クラスタの単語ベクトルはそのクラスタに含まれているページの単語ベクトルを足し合わせたものとする．すなわち，クラスタ $\boldsymbol{c}$ における単語 t の重み $w(t,\boldsymbol{c})$ は以下の式で与えられる．

$$w(t,\boldsymbol{c}) = \sum_{p \in \boldsymbol{c}} w(t,p) \tag{3.4}$$

ただし，この計算式での「クラスタに含まれているページ」は基本ページを指し，クラスタに新しいページが追加されてもそのクラスタの単語ベクトルを更新しないことにする．これは，新しく追加されたページがそのクラスタとほとんど関係のないノイズだった場合に受ける悪影響を防ぐとともに，単語ベクトルの再計算による処理時間の増加を避けるためである．

例

単純な例として，5 つのページ $(p_n,\ n = 1,\ldots,5)$ について，5 種類の単語 $(t_m,\ m = 1,\ldots,5)$ の頻度が以下のように与えられている．

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & p_4 & p_5 \end{array} \\ \begin{pmatrix} 3 & 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 & 3 \end{pmatrix} \end{array}$$

また，$p1,\ p2$ がクラスタ $\boldsymbol{c}_1$，$p5$ がクラスタ $\boldsymbol{c}_2$ の基本ページである．この条件の下でクラスタリングを行うとする．

その他クラスタを含めて $N_c = 3$ である．したがって cf は

$$cf(\boldsymbol{t}) = \begin{pmatrix} 3 \\ 1 \\ 2 \\ 3 \\ 2 \end{pmatrix}$$

となるから，ICF 値は

$$icf(\boldsymbol{t}) = \begin{pmatrix} \ln\left(\frac{3}{3}+1\right) \\ \ln\left(\frac{3}{1}+1\right) \\ \ln\left(\frac{3}{2}+1\right) \\ \ln\left(\frac{3}{3}+1\right) \\ \ln\left(\frac{3}{2}+1\right) \end{pmatrix} \simeq \begin{pmatrix} 0.693 \\ 1.386 \\ 0.916 \\ 0.693 \\ 0.916 \end{pmatrix}.$$

重み TF-ICF 値を計算すると，

$$w(\boldsymbol{t},\boldsymbol{p}) = tficf(\boldsymbol{t},\boldsymbol{p})$$

$$= \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & p_4 & p_5 \end{array} \\ \begin{pmatrix} \frac{3}{4}\times 0.693 & \frac{1}{2}\times 0.693 & \frac{2}{4}\times 0.693 & \frac{0}{4}\times 0.693 & \frac{1}{5}\times 0.693 \\ \frac{0}{4}\times 1.386 & \frac{0}{2}\times 1.386 & \frac{1}{4}\times 1.386 & \frac{0}{4}\times 1.386 & \frac{0}{5}\times 1.386 \\ \frac{1}{4}\times 0.916 & \frac{0}{2}\times 0.916 & \frac{1}{4}\times 0.916 & \frac{0}{4}\times 0.916 & \frac{0}{5}\times 0.916 \\ \frac{0}{4}\times 0.693 & \frac{1}{2}\times 0.693 & \frac{0}{4}\times 0.693 & \frac{2}{4}\times 0.693 & \frac{1}{5}\times 0.693 \\ \frac{0}{4}\times 0.916 & \frac{0}{2}\times 0.916 & \frac{0}{4}\times 0.916 & \frac{2}{4}\times 0.916 & \frac{3}{5}\times 0.916 \end{pmatrix} \end{array}$$

$$\simeq \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & p_4 & p_5 \end{array} \\ \begin{pmatrix} 0.520 & 0.347 & 0.347 & 0 & 0.139 \\ 0 & 0 & 0.347 & 0 & 0 \\ 0.229 & 0 & 0.229 & 0 & 0 \\ 0 & 0.347 & 0 & 0.347 & 0.139 \\ 0 & 0 & 0 & 0.347 & 0.550 \end{pmatrix} \end{array}.$$

したがって，基本クラスタの単語ベクトルは，

$$
w(\boldsymbol{t},\boldsymbol{c}) = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{array}
\begin{array}{c}
\begin{array}{cc} \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{array} \\
\begin{pmatrix}
0.520 + 0.347 & 0.139 \\
0 + 0 & 0 \\
0.229 + 0 & 0 \\
0 + 0.347 & 0.139 \\
0 + 0 & 0.550
\end{pmatrix}
\end{array}
$$

$$
= \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{array}
\begin{array}{c}
\begin{array}{cc} \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{array} \\
\begin{pmatrix}
0.866 & 0.139 \\
0 & 0 \\
0.229 & 0 \\
0.347 & 0.139 \\
0 & 0.550
\end{pmatrix}
\end{array}.
$$

$p_3,\ p_4,\ \boldsymbol{c}_1,\ \boldsymbol{c}_2$ のユークリッドノルムは

$$
\begin{aligned}
||p_3|| &= \sqrt{0.347^2 + 0.347^2 + 0.229^2} \\
&\simeq 0.542, \\
||p_4|| &= \sqrt{0.347^2 + 0.458^2} \\
&\simeq 0.575, \\
||\boldsymbol{c}_1|| &= \sqrt{0.866^2 + 0.229^2 + 0.347^2} \\
&\simeq 0.961, \\
||\boldsymbol{c}_2|| &= \sqrt{0.139^2 + 0.139^2 + 0.550^2} \\
&\simeq 0.584.
\end{aligned}
$$

ゆえに，類似度は

$$
\begin{aligned}
s(\boldsymbol{p}, \boldsymbol{c}) &= \begin{matrix} \\ p_3 \\ p_4 \end{matrix} \begin{matrix} \boldsymbol{c}_1 \qquad\qquad \boldsymbol{c}_2 \end{matrix} \\
&\quad \begin{pmatrix} \dfrac{p_3 \cdot \boldsymbol{c}_1}{||p_3|| \times ||\boldsymbol{c}_1||} & \dfrac{p_3 \cdot \boldsymbol{c}_2}{||p_3|| \times ||\boldsymbol{c}_2||} \\ \dfrac{p_4 \cdot \boldsymbol{c}_1}{||p_4|| \times ||\boldsymbol{c}_1||} & \dfrac{p_4 \cdot \boldsymbol{c}_2}{||p_4|| \times ||\boldsymbol{c}_2||} \end{pmatrix} \\
&= \begin{matrix} \\ p_3 \\ p_4 \end{matrix} \begin{matrix} \boldsymbol{c}_1 \qquad\qquad\qquad\qquad\qquad\qquad \boldsymbol{c}_2 \end{matrix} \\
&\quad \begin{pmatrix} \dfrac{0.347 \times 0.866 + 0.229 \times 0.229}{0.542 \times 0.961} & \dfrac{0.347 \times 0.139}{0.542 \times 0.584} \\ \dfrac{0.347 \times 0.347}{0.575 \times 0.961} & \dfrac{0.347 \times 0.139 + 0.458 \times 0.550}{0.575 \times 0.584} \end{pmatrix} \\
&= \begin{matrix} \\ p_3 \\ p_4 \end{matrix} \begin{matrix} \boldsymbol{c}_1 \qquad \boldsymbol{c}_2 \end{matrix} \\
&\quad \begin{pmatrix} 0.678 & 0.152 \\ 0.218 & 0.894 \end{pmatrix}.
\end{aligned}
$$

類似度が最大である組み合わせは p_4 と $\boldsymbol{c}_2$ であり，その値は $0.894 > 0.4$ であるから，p_4 を $\boldsymbol{c}_2$ に追加する．

単語ベクトルはクラスタの追加によって変化しないため，残った p_3 とクラスタ $\boldsymbol{c}_1$，$\boldsymbol{c}_2$ の類似度も変化せず

$$
s(p_3, \boldsymbol{c}) = \begin{matrix} \\ p_3 \end{matrix} \begin{matrix} \boldsymbol{c}_1 \qquad \boldsymbol{c}_2 \end{matrix} \begin{pmatrix} 0.678 & 0.152 \end{pmatrix}.
$$

類似度が最大であるのは $\boldsymbol{c}_1$ との組み合わせであり，その値は $0.678 > 0.4$ であるから，p_3 を $\boldsymbol{c}_1$ に追加する．この時点でその他のページがなくなったのでクラスタリングを終了する．

最終的には p_1，p_2，p_3 が $\boldsymbol{c}_1$ クラスタに属し，p_4，p_5 が $\boldsymbol{c}_2$ クラスタに属すことになる．

第4章　評価実験

4.1　実験方法

本研究の目的はリンク集の自動生成であり，それにはユーザがどのようなテーマを与えても適切なリンク集を構築する必要がある．また，テーマに曖昧性がある場合には適切なクラスタを構築しなければならない．そこで，本実験では表 4.1 の 5 組のテーマについて 3 節で述べた手法を適用し，評価を行う．

表 4.1: 評価実験用のテーマ

テーマ番号	テーマ (キーワード)		
1	松井		
2	石川	テレビ	番組表
3	perl	リファレンス	
4	地図	日本	
5	野球		

評価基準は以下の通りである．

- リンク集が正確に検出できたかどうか
- 基本クラスタと基本ページ集合が適切かどうか
- クラスタへのページ追加処理でキーワードの曖昧性が正しく考慮されているかどうか
- 候補ページを追加した効果があるかどうか

これらの評価の際には，対象となるページ集合の中からいくつかのサンプルを取り出して人手で正解を付け，システムの結果と照合する．評価基準の詳細は実験結果と共に次節で述べる．

4.2 評価基準と実験結果

4.2.1 リンク集の検出

ここでは 3.3 節で述べたリンク集の検出手法について評価する．

各ページについてリンク集部分が正しく検出されているかどうかの評価を行った．本実験では 1 組のテーマにつき，評価用のサンプルとしてシステムがリンク集部分を検出したページ 15 件とリンク集部分を検出しなかったページ 15 件を選んだ．これらのサンプルについて，リンク集単位でエラー率 (Error rate: E)，適合率 (Precision: P)，再現率 (Recall: R) を求めた．ページ単位ではなくリンク集単位で各評価尺度を計算したのは 1 つのページにリンク集部分が複数存在することがあるためである．それぞれの評価尺度は以下の式で算出した．

$$E = \frac{(\text{リンク集を誤検出したページの数})}{(\text{リンク集をもたないページの数})} \tag{4.1}$$

$$P = \frac{(\text{正しく検出したリンク集の数})}{(\text{検出したリンク集の数})} \tag{4.2}$$

$$R = \frac{(\text{正しく検出したリンク集の数})}{(\text{実際のリンク集の数})} \tag{4.3}$$

結果を表 4.2 に示す．

表 4.2: リンク集の検出結果

テーマ番号	エラー率	適合率	再現率
1	8.3% (1 / 12)	95.5% (21 / 22)	53.8% (21 / 39)
2	42.1% (8 / 19)	52.9% (9 / 17)	34.6% (9 / 26)
3	37.5% (9 / 24)	65.4% (17 / 26)	68.0% (17 / 25)
4	34.8% (8 / 23)	50.0% (8 / 16)	57.1% (8 / 14)
5	31.3% (5 / 16)	66.7% (10 / 15)	71.4% (10 / 14)
全体	33.0% (31 / 94)	67.7% (65 / 96)	55.1% (65 / 118)

テーマによって偏りが見られるが，全体としては適合率の方が再現率より高かった．適合率が低い場合は本来リンク集でないものまでリンク集と判断しているということであり，追加する候補ページにノイズが混入する可能性が高まる．しかし，追加する候補ページにはキーワードが全て含まれるかどうかのチェックを行うため，リンク集でないものを誤検出してもテーマに関連のないページが追加されることは少ないと考えられる．対して，再現率が低い場合は本来リンク集であるものを検出できていないということであるから，追加する候補ページが少なくなることに加えリンク集のみからなるページを正確に削除できなくなる可能性が高まる．ページ数が増えるとクラスタリングの処理時間も増え

るが，中には有用なページも含まれているため，追加する候補ページが少なくするのは得策ではないといえる．また，テーマに関連があるリンク先ページは候補に追加されるのだから，リンク集のみからなるページを残しておくメリットは特にない．したがって，本研究においてはエラー率や適合率を多少犠牲にしても再現率を上げる方が有効と考えられ，本実験の結果はあまり良い結果であるとはいえない．リンク集に関しては再現率を向上させることが当面の課題といえる．

適合率を下げた大きな要因は内部リンクを外部リンクと誤判別したというものである．理論上はリンク元ページの URL とリンク先ページの URL について，独自ドメインをもつページならばドメイン部分を，そうでないページはユーザディレクトリまでの部分を比較すれば正しく判別できると考えられる．しかし，リンク元ページが独自ドメインをもつかどうかを判断するための基準がなく，どのページに対してもドメインとその直下のディレクトリで判断しているのが誤判別の原因となっている．例えば，JAIST の知識科学研究科のページ (“http://www.jaist.ac.jp/ks/index.html”)，情報科学研究科のページ (“http://www.jaist.ac.jp/is/index-jp.html”)，材料科学研究科のページ (“http://www.jaist.ac.jp/ms/index.html”) は全て JAIST 内のページである．ところが，ドメインとその直下のディレクトリで判断すると，“http://www.jaist.ac.jp/ks/”，“http://www.jaist.ac.jp/is/”，“http://www.jaist.ac.jp/ms/” と，ディレクトリ部分が異なるため外部サイトと判断されてしまう．

再現率が低い要因はリンク集のパターンが少ないことによるところが大きい．製作者によってウェブページの記述方法が大きく異なるため，リンク集のパターンも 3.3 節で述べたものだけでは対応できないものが見られた．その他のリンク集パターンとして見られたものを以下に 3 例示す．

- アンカーの羅列 (改行を含まない)
- 定義リスト (dt 要素内にリンクを含み，dd 要素がリンク先の説明を表す)
- リンクが列方向に並んだ表形式

これらのパターン等を加えることで再現率は向上すると考えられる．しかし，適合率が低下するだけでなく，リンク集のみからならないページ誤ってを削除する危険性も高まるため注意が必要である．

4.2.2 基本クラスタ

ここでは 3.4.1 項で述べた手法で作成された基本クラスタについて評価する．

作成された基本クラスタが適切かどうか，また各基本ページがリンク集に掲載すべきページとして適切かどうかという基準で評価を行った．後者については以下の 2 点を基準とした．

- クラスタと関連性の高いページかどうか

- リンク集のみからなるページでないか

クラスタとの関連性はクラスタと本文の主要な内容がマッチしているかどうかで判断した．したがって，クラスタ名の元となった語が本文の一部のみに現れたものやナビゲーション部などの本文でない部分に現れたものは不正解とみなした．

基本ページのサンプル数は初期候補ページ 15 件と，3.2.3 項の手法で追加した候補ページ 15 件の計 30 ページとし，どちらかが 15 件に満たない場合はもう一方で充填した．また，基本ページ数が 30 件に満たないクラスタは全ての基本ページをサンプルとした．

まず，作成された基本クラスタの評価を行う．作成されたクラスタを表 4.3 に示す．

表 4.3: 基本クラスタ

テーマ番号	クラスタ名	基本ページ数
1	松井 秀喜	102
	松井 稼	23
	松井 雄飛	22
2	石川 県	71
	石川 テレビ	79
	ケーブル テレビ	34
	テレビ 番組表	32
	テレビ 番組	53
	テレビ 番組表	91
	週間 番組表	28
3	perl5 リファレンス	27
	ポケット リファレンス	28
	perl リファレンス	39
4	日本 地図	78
	日本 地図	80
	日本 全国	25
5	高校 野球	99
	プロ 野球 ニュース	73
	プロ 野球	148
	高校 野球 部	56
全体		1188

表中の下線部はキーワードを示す．また，表中の「松井稼」クラスタは「松井稼頭央」クラスタである．茶筌では「稼頭央」が名前だと認識されず「稼」だけで 1 つの形態素と扱われてしまうが，ニュースサイトなどでは「松井稼」と略されることもあるため，むしろ基本ページを増やす働きをしていると思われる．

表 4.3 を見ると，テーマ 1 では 3 人の「松井」に関するクラスタが生成され，人物以外のクラスタが作られていないものの比較的成功した例だといえる．対して，テーマ 2 では「テレビ番組表」と「テレビ番組」のように似た名前のクラスタが作られた．テーマ 3 の「perl5 リファレンス」と「perl リファレンス」なども同様に似た名前のクラスタとなっている．これは，3.4 節で述べたアルゴリズムが「キーワード前後の名詞が 1 文字でも違えば異なる」と判断しているためである．

例えば，テレビ番組表とテレビ番組では考え方にもよるが意味が異なるように思われる．対して，perl5 と perl では概念的に perl5 $\subset$ perl であることは疑いない．したがって，場合によっては複数のクラスタを統合することでクラスタを洗練することができると考えられる．

テーマ 3 において，キーワード「perl」をクラスタの基準としたものではクラスタが作成されなかった．これは，キーワードの前後両方に名詞が存在する場合が極端に少ないときに起こる．タイプミス等を除けば，perl という語は通常プログラミング言語の Perl を指すので曖昧性は存在しないといえる．その考えに基づけばクラスタによって細分化されなかったのは成功と考えてよい．

基本ページの適合率

次に，作成された基本クラスタに含まれる基本ページが適切かどうかを評価する．基本ページはシステムが選んだものであるから，評価尺度には適合率を用いる．

$$P = \frac{\text{(掲載に適したページの数)}}{\text{(基本ページのサンプル数)}} \tag{4.4}$$

各クラスタおよび全サンプルでの結果を表 4.4 に示す．

基本ページの適合率は個々のクラスタ間で大きく異なっていたが，全体的に見ると一般的な単語を含むクラスタはそうでないクラスタに比べて適合率は低い傾向にあるといえる．一般的な単語は直接の関連がないページにも出現する割合が多く，結果的にノイズとなりやすいことが理由として考えられる．例えば，テーマ 1 の「松井」では，おそらく「松井雄飛」より知名度が高い＝一般的と考えられる「松井秀喜」や「松井稼 (頭央)」のクラスタの方が適合率が低かった．また，他のテーマに比べて全体的に適合率が低かったテーマ 2 では「テレビ」というキーワード自身が一般的な語である．

テーマ 1 では前述したように「松井雄飛」クラスタの適合率が他の 2 つに比べて高かった.「松井秀喜」クラスタや「松井稼 (頭央)」クラスタの適合率が低い理由は，30 件の中に直接彼らに関するページではないが名前が挙げられているというページが存在したためである．例えば，別の野球選手に関するページ中で引き合いに名前を出され，クラスタの基本ページとなる条件が整ってしまったページが見られた．クラスタが人物の場合，他のページでの取り上げられやすさと適合率に負の相関があると考えられる．

テーマ 2 では，リンク集に掲載すべきではないのに基本クラスタに含まれていたページが多く見られた．これは，キーワードが本文の主要な部分には含まれておらず，リンク集

表 4.4: 基本クラスタの適合率

テーマ番号	クラスタ名	適合率
1	松井 秀喜	53.3% (16 / 30)
	松井 稼	56.5% (13 / 23)
	松井 雄飛	81.8% (18 / 22)
2	石川 県	30.0% (9 / 30)
	石川 テレビ	6.7% (2 / 30)
	ケーブル テレビ	33.3% (10 / 30)
	テレビ 番組表	20.0% (6 / 30)
	テレビ 番組	50.0% (15 / 30)
	テレビ 番組表	23.3% (7 / 30)
	週間 番組表	3.3% (1 / 28)
3	perl5 リファレンス	92.6% (25 / 27)
	ポケット リファレンス	28.6% (8 / 28)
	perl リファレンス	23.3% (7 / 30)
4	日本 地図	50.0% (15 / 30)
	日本 地図	63.3% (19 / 30)
	日本 全国	24.0% (6 / 25)
5	高校 野球	73.3% (22 / 30)
	プロ 野球 ニュース	13.3% (4 / 30)
	プロ 野球	46.7% (14 / 30)
	高校 野球 部	90.0% (27 / 30)
全体		42.6% (244 / 573)

部分のアンカーテキストに含まれているページが多かったことが原因である．例えば，本文では「石川」については全く触れていないのに，リンク集部分に石川テレビへのリンクがあったため「石川テレビ」クラスタに属したページがある．キーワードがアンカーテキストに含まれているページはこのテーマに限らず存在したが，このテーマの場合はテレビ局や番組表へのリンクがあるページが多かったため，特に影響が大きかった．アンカーテキストとなっているものを除外してクラスタを作成した方が適合率が上がる可能性があるため，今後検討したい．

テーマ 3 では「perl5 リファレンス」クラスタの適合率が非常に高かった．一方,「perl リファレンス」クラスタの適合率はさほど良くはなかった.「perl リファレンス」という語はアンカーテキストなどのノイズとして現れることも多かったが,「perl5 リファレンス」ではノイズとなることが極端に少なかった.「perl5 リファレンス」クラスタは対象が非常に限定されているため，前述したクラスタの統合により「perl リファレンス」のノイズを抑えた方が全体としては良いのではないかと思われる．

テーマ 4 では「日本地図」クラスタが 2 種類作成されたが，適合率は 2 つの間で差が見られた．これは一方が「日本」を基準に，もう一方が「地図」を基準にしてクラスタが作成されたことによる.「日本」基準の場合は直前の名詞が「[名詞なし]」のページが集まっているが,「地図」基準の場合は日本の直前の名詞を考慮してはいない．例えば「バカ日本地図」に関するページは前者の基本ページには含まれないが後者の基本ページには含まれるといったことになる．これらの差が適合率の違いに現れたと考えられる．

テーマ 5 では高校野球に関するクラスタ 2 つは良い結果が得られているが，プロ野球に関するクラスタ 2 つはあまり良い結果が得られなかった．特に「プロ野球ニュース」クラスタはテーマ 2 のクラスタと同様にリンク集部分のアンカーテキストが悪影響を及ぼしたと見られるものが多かった．

各テーマに共通して見られたのは，リンク集のみからなるページが完全に削除されていなかったというものである．3.2.4 項で述べたように，リンク集のみからなるページはリンク集に掲載すべきページとしてはふさわしくなく，適合率を下げる要因となる．これに関してはリンク集の検出精度を上げることで対応できると考えられる．

候補ページの追加による効果

候補ページを追加したことによってより良いリンク集が作成できたかどうかを調べるため，サンプルを初期候補ページ 15 件のみと 3.2.3 項の手法で追加した候補ページ 15 件のみに分けた．この場合の結果を表 4.5 に示す．

サンプルが少なすぎるものはあまり参考にはならないが，テーマ 1 やテーマ 2 などの結果を見る限り追加した候補ページのみでの結果は初期候補ページのみでの結果に比べて適合率が低いといえる．しかし,「テレビ番組」クラスタや「高校野球」クラスタのように特に差が見られないもの,「perl5 リファレンス」クラスタや「高校野球部」クラスタのように十分良いといえるものも存在した．少なくともこれらのクラスタでは Goo の検索

表 4.5: 基本クラスタの適合率 (候補ページ分割時)

テーマ番号	クラスタ名	適合率 (初期候補)		適合率 (追加した候補)	
1	松井 秀喜	93.3%	(14 / 15)	13.3%	(2 / 15)
	松井 稼	91.7%	(11 / 12)	18.2%	(2 / 11)
	松井 雄飛	100.0%	(4 / 4)	77.8%	(14 / 18)
2	石川 県	60.0%	(9 / 15)	0.0%	(0 / 15)
	石川 テレビ	13.3%	(2 / 15)	0.0%	(0 / 15)
	ケーブル テレビ	57.1%	(8 / 14)	12.5%	(2 / 16)
	テレビ 番組表	26.7%	(4 / 15)	13.3%	(2 / 15)
	テレビ 番組	46.7%	(7 / 15)	53.3%	(8 / 15)
	テレビ 番組表	40.0%	(6 / 15)	6.7%	(1 / 15)
	週間 番組表	6.7%	(1 / 15)	0.0%	(0 / 13)
3	perl5 リファレンス	100.0%	(1 / 1)	92.3%	(24 / 26)
	ポケット リファレンス	29.6%	(8 / 27)	0.0%	(0 / 1)
	perl リファレンス	24.1%	(7 / 29)	0.0%	(0 / 1)
4	日本 地図	53.6%	(15 / 28)	0.0%	(0 / 2)
	日本 地図	72.0%	(18 / 25)	20.0%	(1 / 5)
	日本 全国	26.1%	(6 / 23)	0.0%	(0 / 2)
5	高校 野球	73.3%	(11 / 15)	73.3%	(11 / 15)
	プロ 野球 ニュース	100.0%	(4 / 4)	0.0%	(0 / 26)
	プロ 野球	66.7%	(10 / 15)	26.7%	(4 / 15)
	高校 野球 部	80.0%	(8 / 10)	95.0%	(19 / 20)
全体		49.4%	(154 / 312)	34.5%	(90 / 261)

で見つけられなかったページをリンク集に加えることができ，候補ページを追加した効果があったといえる．

また，前述したように不正解となったものはアンカーテキストや削除されていないリンク集ページの影響によるものが多い．したがってこれらの点を改善すれば，逆効果となったクラスタについても候補ページを追加したことによる効果が見られるようになる可能性はある．現時点では，追加した候補ページは初期候補ページの足を引っ張る結果となっているが，問題点を改善してからあらためて評価する必要がある．

4.2.3 クラスタへのページ追加

ここでは 3.4.2 項で述べた基本クラスタへのページの追加手法について評価する．

3.4.2 項の処理で基本クラスタに追加されたページについて，キーワードの意味と合うクラスタに追加されているかどうかという基準で評価を行った．なお，4.2.2 項ではリンク集に掲載すべきページとして適切かどうかで評価したが，本項ではキーワードの意味のみから評価するため，掲載ページとして適切かどうかは考慮しない．これはあくまで純粋なクラスタリングの性能を評価する目的のためである．同じキーワードが複数回出現するページに関しては，もっとも多く使われている意味で判断した．

サンプルはクラスタに追加されたページから初期候補ページ 15 件と既存のリンク集を辿ることで追加した候補ページ 15 件の計 30 ページを選んだ．どちらかが 15 件に満たない場合はもう一方で充填し，クラスタに追加されたページの数が 30 件に満たない場合はそれらのページ全てをサンプルとした．評価尺度は適合率である．

$$P = \frac{(\text{キーワードの意味と合うクラスタに追加されたページの数})}{(\text{クラスタに追加したページのサンプル数})} \tag{4.5}$$

各クラスタおよび全サンプルでの結果を表 4.6 に示す．

基本ページの適合率と同様に，クラスタに追加されたページの適合率も個々のクラスタ間で大きく異なっていた．また，追加されたページの少ない，もしくはページの追加されていないクラスタも多い．

適合率の低いクラスタを見ると，基本ページでの適合率も良くなかったものが多い．基本クラスタでノイズとなっていたページはアンカーテキストや削除されなかったリンク集ページが多く，基本的にキーワードが本文中に出現していない．その場合，キーワードの周りの単語はあまり有用ではない可能性が高く，結果的にクラスタ側の単語ベクトルが無意味なものになっており，悪影響を及ぼしていると思われる．

適合率が 5 位以内のクラスタは追加されたページの 2/3 以上が正解であった．これらに関してはクラスタリングが比較的うまくいったといえる．しかし,「松井秀喜」クラスタについては,「松井」というと多くの場合松井秀喜を指すため，偶然当たっただけという可能性がある．実際，表 4.4 の基本ページ数を見ると，掲載ページとしての良し悪しはともかく，他のクラスタ 2 つの 5 倍近いページが「松井秀喜」クラスタに割り当てられている.「テレビ番組表」クラスタに関しても同様に,「番組表」といったらテレビ番組表を指

表 4.6: 追加されたページの適合率

テーマ番号	クラスタ名	追加ページ数	適合率
1	松井 秀喜	24	87.5% (21 / 24)
	松井 稼	0	-
	松井 雄飛	0	-
2	石川 県	3	33.3% (1 / 3)
	石川 テレビ	150	10.0% (3 / 30)
	ケーブル テレビ	106	3.3% (1 / 30)
	テレビ 番組表	20	30.0% (6 / 20)
	テレビ 番組	32	40.0% (12 / 30)
	テレビ 番組表	85	76.7% (23 / 30)
	週間 番組表	1	0.0% (0 / 1)
3	perl5 リファレンス	0	-
	ポケット リファレンス	5	20.0% (1 / 5)
	perl リファレンス	88	66.7% (20 / 30)
4	日本 地図	102	50.0% (15 / 30)
	日本 地図	41	40.0% (12 / 30)
	日本 全国	1	0.0% (0 / 1)
5	高校 野球	70	80.0% (24 / 30)
	プロ 野球 ニュース	1	0.0% (0 / 1)
	プロ 野球	119	30.0% (9 / 30)
	高校 野球 部	77	93.3% (28 / 30)
全体		925	49.6% (176 / 355)

すことが多いために適合率が高くなった可能性がある．したがって，この 2 つに関しては十分な調査が必要だろう．

候補ページの追加による効果

4.2.2 項での評価と同様に候補ページの追加による効果を調べるため，サンプルを初期候補ページ 15 件のみと 3.2.3 項の手法で追加した候補ページ 15 件のみに分けた．このときの結果を表 4.7 に示す．

表 4.7: 追加されたページの適合率 (候補ページ分割時)

テーマ番号	クラスタ名	適合率 (初期候補)	適合率 (追加した候補)
1	松井 秀喜	85.0% (17 / 20)	100.0% (4 / 4)
	松井 稼	-	-
	松井 雄飛	-	-
2	石川 県	50.0% (1 / 2)	0.0% (0 / 1)
	石川 テレビ	13.3% (2 / 15)	6.7% (1 / 15)
	ケーブル テレビ	6.7% (1 / 15)	0.0% (0 / 15)
	テレビ 番組表	50.0% (4 / 8)	16.7% (2 / 12)
	テレビ 番組	36.8% (7 / 19)	45.5% (5 / 11)
	テレビ 番組表	66.7% (10 / 15)	86.7% (13 / 15)
	週間 番組表	0.0% (0 / 1)	-
3	perl5 リファレンス	-	-
	ポケット リファレンス	20.0% (1 / 5)	-
	perl リファレンス	63.6% (14 / 22)	75.0% (6 / 8)
4	日本 地図	70.6% (12 / 17)	23.1% (3 / 13)
	日本 地図	46.2% (12 / 26)	0.0% (0 / 4)
	日本 全国	-	0.0% (0 / 1)
5	高校 野球	60.0% (9 / 15)	100.0% (15 / 15)
	プロ 野球 ニュース	-	0.0% (0 / 1)
	プロ 野球	33.3% (5 / 15)	26.7% (4 / 15)
	高校 野球 部	50.0% (1 / 2)	96.4% (27 / 28)
全体		48.7% (96 / 197)	50.6% (80 / 158)

ページ 30 件での評価で適合率が高かったクラスタに関しては追加した候補ページのみの適合率が，逆に 30 件での評価で適合率が低かったクラスタに関しては初期候補ページのみの適合率が高い傾向にあった．前述の通り，30 件の評価で適合率が低かったものはクラスタの単語ベクトル自体に問題のある可能性が高く，追加した候補ページが無意味だ

と断定はできない．上位の適合率の高さもあり，全体的には候補ページの追加処理は効果的であると考えられる．

第5章　結論

5.1　まとめ

本研究では関連リンク集の自動生成を目的とし，掲載するウェブページの取得・選別を試みた．その際，掲載するページを追加したり不要なページを削除するためにリンク集の検出を行った．また，ユーザの求める適切なリンク集を構築するため，キーワードの曖昧性を考慮したクラスタリングを行った．

評価実験の結果，リンク集の検出については再現率が 55%程度であった．生成されるクラスタに関しては，あるテーマでは適切なクラスタが生成されたが，別のテーマではあまり適切でないクラスタが生成されてしまった．また，リンク集に掲載すべきでないページまで掲載してしまうなど，クラスタの基本ページの適合率は 5 割程度であった．クラスタリングの評価では，クラスタに追加されたページの中でクラスタとキーワードの意味が合致していたものの割合は 42.6%程度であった．候補ページの追加処理については，ノイズの混入も多かったが，効果のあったクラスタもいくつか見られた．

5.2　今後の課題

今回，リンク集の検出において定義したパターンが十分ではなかった．そこで，パターンを増やし，再現率を向上させる必要がある．その結果，今回の評価実験でクラスタリングの際に残ってしまったリンク集ページを除去でき，クラスタリングの精度向上にもつながると期待される．新たに増やすパターンは 4.2.1 項で述べた 3 種類だけでなく，別のパターンも考えられないか調査・検討する必要がある．

次に，基本クラスタの生成で，与えたテーマによっては似た名前のクラスタが生成された．このとき，一方のクラスタがもう一方のクラスタの部分集合となっているときはそれらのクラスタを統合することでクラスタが洗練される可能性がある．今後，ユーザの利便性向上のためにも，どのような場合にクラスタを統合すればクラスタが洗練されるかを検討すべきである．また，クラスタの統合によって基本ページやクラスタに追加されたページの精度がどう変わるか調べる必要がある．

同じく基本クラスタを生成したときに，その初期のクラスタを構成する基本ページにノイズが多く含まれていた．ノイズには，リンク集部分のアンカーテキストなどページの主要部と直接関係のない部分にキーワードが出現したものがあり，不適切なクラスタに属してしまったページが存在した．そこで，3.4.1 項の名詞の抽出ステップを改良し，ページ

内で本文と直接関係のある部分に出現するキーワードとその前後の名詞だけを取り出すといったことを検討すべきである．

本研究では素性としてキーワードの前後 50 単語以内の自立語を選択した．しかし，これが最良であるとはいえない．そこで，クラスタリングの単語ベクトルの素性選択において，より優れた方法がないか検討する必要がある．また，重み付けの式として TF-ICF 値を提案したが，よりクラスタを特徴付ける重み付けがないか検討すべきである．

謝辞

本研究を進めるにあたり，熱心なご指導を賜りました白井清昭助教授に心から感謝いたします．また，多くのご教示を賜りました島津明教授に心から感謝いたします．多くのご助言を頂きました山田寛康助手，中村誠助手に深く感謝いたします．自然言語処理学講座の皆様には，貴重なご意見，ご支援を頂きましたことを感謝いたします．

参考文献

[1] 平野健児, 第三者による解説・評価を含む Web 関連リンク集の自動生成. Master's thesis, 北陸先端科学技術大学院大学, 2004.

[2] Satoshi Sato, Madoka Sato: Automatic Generation of Web Directories for Specific Categories. *AAAI Workshop on Intelligent Information Systems*, Orlando, July, 18-19, 1999.

[3] Oren Zamir, Oren Etzioni: Web Document Clustering: A Feasibility Demonstration. *SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp.46-54, August 24-28, 1998.

[4] 金子大輔, 高山毅, 池田哲夫, 長内亘: Gooots–検索目的に沿ってカテゴリ名を取捨選択してクラスタリングする検索エンジン. 情報処理学会第 67 回全国大会, 2005.

[5] Vivísimo, `http://vivisimo.com/`

[6] ChaSen's Wiki, `http://chasen.naist.jp/hiki/ChaSen/`

[7] syger.com - The English language stop-words,
`http://www.syger.com/jsc/docs/stopwords/english.htm`

[8] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney: Impact of Similarity Measures on Web-page Clustering. *AAAI 2000: Workshop of Artificial Intelligence for Web Search*, pp.58-64, July, 2000.