修 士 論 文

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

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

平野 健児

2004 年 3 月

修 士 論 文

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

指導教官　白井 清昭

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

210076 平野 健児

提出年月: 2004年2月

概要

本論文では，Web 探索を行うユーザが欲しい情報を簡単に Web から獲得するために，ユーザが要求するキーワードに関連した Web ページを収集し，関連リンク集を自動生成する事を目的とする．また，自動収集された Web ページを提示するだけでは，ユーザが一つ一つ Web ページを閲覧していかなければ，有用な情報がどこにあるのかわからない．そこで，各 Web ページに関する解説も提示する．本論文では，関連リンク集に掲載するにページ（以下，対象ページ）に関する解説は Web ページ自身から取り出すのではなく，その Web ページにリンクをはっているページ（以下，参照ページ）から取り出す．これにより，Web ページ自身からは得られない第三者による記述をユーザに提示することができる．また，Web ページに関する記述を説明・評価などに分類し，ユーザにわかりやすく提示する．また，関連リンク集をユーザに提示する際，掲載するページの収集・選別も重要であるが，本論文ではリンク先ページの説明・評価の記述に重点を置く．本システムは以下のステップにより構成される．

1. ユーザーによるキーワード入力
2. キーワードに関するページの URL の収集
3. 参照ページからの対象ページに関する記述の抽出
4. 得られた記述の説明文・評価文などへの分類
5. 関連リンク集の出力

最初に，ステップ１～３を行う．具体的には，まずユーザーがキーワードを入力する．キーワードを Goo にかけ，キーワードに関する対象ページの URL を収集する．次に，各々の対象ページの参照ページを収集する．収集の方法は Goo のリンク先 URL 検索を用いて収集する．

次にステップ４について述べる，本論文では，HTML タグ及びサイト名を手掛かりとした二つの方法で Web ページに関する情報を抽出する．

HTML タグを用いた方法は，リストタグ，br タグ，テーブルタグ等を手掛かりに，アンカー周辺にある Web ページの情報を抽出する．リストタグの場合，該当アンカーの前の li タグから，該当アンカーの後ろの li タグまでを Web ページに関する情報として抽出する．br タグは「アンカー＋文字列＋ br」が 3 回以上並んでいる場合，この文字列を Web ページに関する情報として抽出する．テーブルタグの場合，アンカーを含むセルの右側のセルにアンカー以外の文字列が記述されていれば，その文字列を参照箇所として取り出す．また，テーブルの同じ列にアンカーと文字列が交互にある場合，該当アンカーの下のセルにある文字列を Web ページに関する情報として抽出する．

サイト名を手掛かりとした方法は，まず，対象ページのサイト名を特定する．具体的には，サイト名として対象ページのアンカーの文字列を抽出する．ただし，長い文字列は，

サイト名として抽出しない．また，サイト名を示す文字列は一つとは限らないので，複数のアンカーからサイト名を複数抽出する．次にサイト名を手掛かりとして Web ページに関する情報を抽出する．具体的には，サイト名の文字列の前にある HTML タグからサイト名の文字列の後ろにある HTML タグまでを Web ページに関する情報として抽出する．

次にステップ５について述べる．本論文では，参照ページから抽出した対象ページに関する記述を「評価：利便」「評価：情報量」「評価：その他」「説明：機能」「説明：記述」のカテゴリに自動的に分類する．カテゴリ「評価：利便」は Web ページに関する利便性，使い勝手といった記述を表す.「評価：情報量」は Web ページの規模，情報量を含む記述を表す.「評価：その他」は情報量と利便性以外の評価を含む記述を表わす.「説明：機能」は Web ページの機能について書かれた記述を表す.「説明：記述」はページの機能以外について書かれたページの説明を表わす．カテゴリの自動分類は,「説明：記述」を除く．それぞれのカテゴリに対してパターンとの人手で作成された．パターンマッチングにより行う．また，４つのカテゴリのいずれのパターンマッチにも失敗したときは，その記述のカテゴリを「説明：記述」とした．

本論文のシステムの評価実験を行った．参照ページから抽出された 467 個の記述に対し，カテゴリの自動分類を行った．カテゴリ [評価:利便] に対する精度は 0.8519，再現率: 0.766 であった．[評価:利便] の場合は，精度は 0.8519，再現率は 0.7692 であった．[評価:その他] の場合，精度は 0.3125，再現率は 0.4167 であった．[説明:機能] は精度 0.6442，再現率 0.7614 であった．[説明:記述] の場合は，精度 0.7797，再現率 0.7302 であった．また，全体の精度は 0.7410，再現率は 0.6617 であった．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　研究の目的・背景

1.1.1　研究の背景

最近、コンピュータの急激な性能向上とPCやインターネットの普及により，Web上で多種多様の情報を誰もが入手できるようになっている．また，誰もがホームページを開設できるようになり，Web上には膨大な情報が蓄積され，どんどん巨大化している．しかし，これらの情報は何かに基づいて整理されたデータベースでもなく，様々な書式・スタイルがあり混沌としている．Web上には有用な情報が多くある．しかし，情報が蓄積されればされるほど，必要とする情報を探し出す事は逆に困難になる．多量の情報の中から，ユーザが必要とする情報を容易にかつ適切に探し出すことを支援する技術が望まれる．

現在の情報探索の方法は，主にYahoo，Goo，Googleといった検索エンジンを用いて，必要とする情報が記述されているWebページを探し出す方法である．検索エンジンを用いた場合，検索結果にWebページに関する記述が書かれている場合が多く見られる．例えば，Googleではユーザが検索にかけたキーワードの周辺の何バイトか取り出して表示している．しかし，このような情報はユーザにとって必ずしも適切な情報とは限らない．そのため有用な情報が記述されているかもしれないWebページを，ユーザはとりあえず一つ一つ確認して判断しなければならない．もし，必要とする情報がなければ，また検索結果が提示する画面に戻り，次のページを確認するという動作を繰り返さなければならない．これはユーザの負担が大きいし，時間もかかる．そこで，リンクをたどる前にWebページに関する適切な情報をユーザに提示することができれば，どのWebページがユーザにとって有用なのか判断できるため，Web探索の効率は上昇し，ユーザにかかる負担は軽くなる

また，情報探索のもうひとつの方法として，リンク集からユーザにとって有用なページを探す方法がある．現存するリンク集には作成者のWebページに対する説明や評価が記述されている場合がある．このような記述はユーザが有用なWebページを探し出すための有力な情報源である．しかし，ユーザが要求するテーマにあった関連リンク集がWeb上に存在するとは限らない．また，リンク集のページに関する説明・評価が書かれているとは限らない．そこで，ユーザが要求するキーワードで関連リンク集が自動的に生成され，第三者の説明・評価といったコメントをWebページと一緒に提示できれば，ユーザが何か調べ物をするときに大変便利である．

1.1.2　本研究の目的

本研究では，ユーザが欲しい情報を簡単に Web から獲得するために，ユーザが要求するキーワードに関連した Web ページを収集し，関連リンク集を自動生成する事を目的とする．また，自動収集された Web ページを提示するだけでは，ユーザが一つ一つ Web ページを閲覧していかなければ，有用な情報がどこにあるのかわからない．そこで，各 Web ページに関する解説も提示する．本研究では，Web ページに関する解説は Web ページ自身から取り出すのではなく，その Web ページにリンクをはっているページから取り出す．これにより，Web ページ自身から得られない第三者による記述をユーザに提示することができる．また，Web ページに関する客観的記述を説明・評価などに分類し，ユーザにわかりやすく提示する．リンク集の自動生成をする際に必要な処理は大きく分けて以下の２つがある．

1. 掲載するページの収集，選別
2. リンク先ページの説明の記述

本研究では，２のリンク先ページの説明の記述に焦点を当てる．特に，Web ページに対する説明は Web 上から自動的に獲得する．また，最終的にはユーザが要求するキーワードを入力とし，図 1.1 のような関連リンク集を出力する事を目的とする．

以下，本論文では，リンク集に載せるページを対象ページ，それにリンクをはっているページを参照ページ，参照ページから抽出する対象ページに関する情報を参照箇所と定義する．対象ページと参照ページの関係を図 1.2 に示す．

1.2　本論文の構成

本論文では，２章で参照箇所抽出と評価文の抽出に関する先行研究を紹介する．３章では関連リンク集の自動生成するための手法と Web ページに関する情報の抽出と分類方法について記述する．４章では提案手法の評価実験について記述する．５章では結論と今後の課題を述べる．

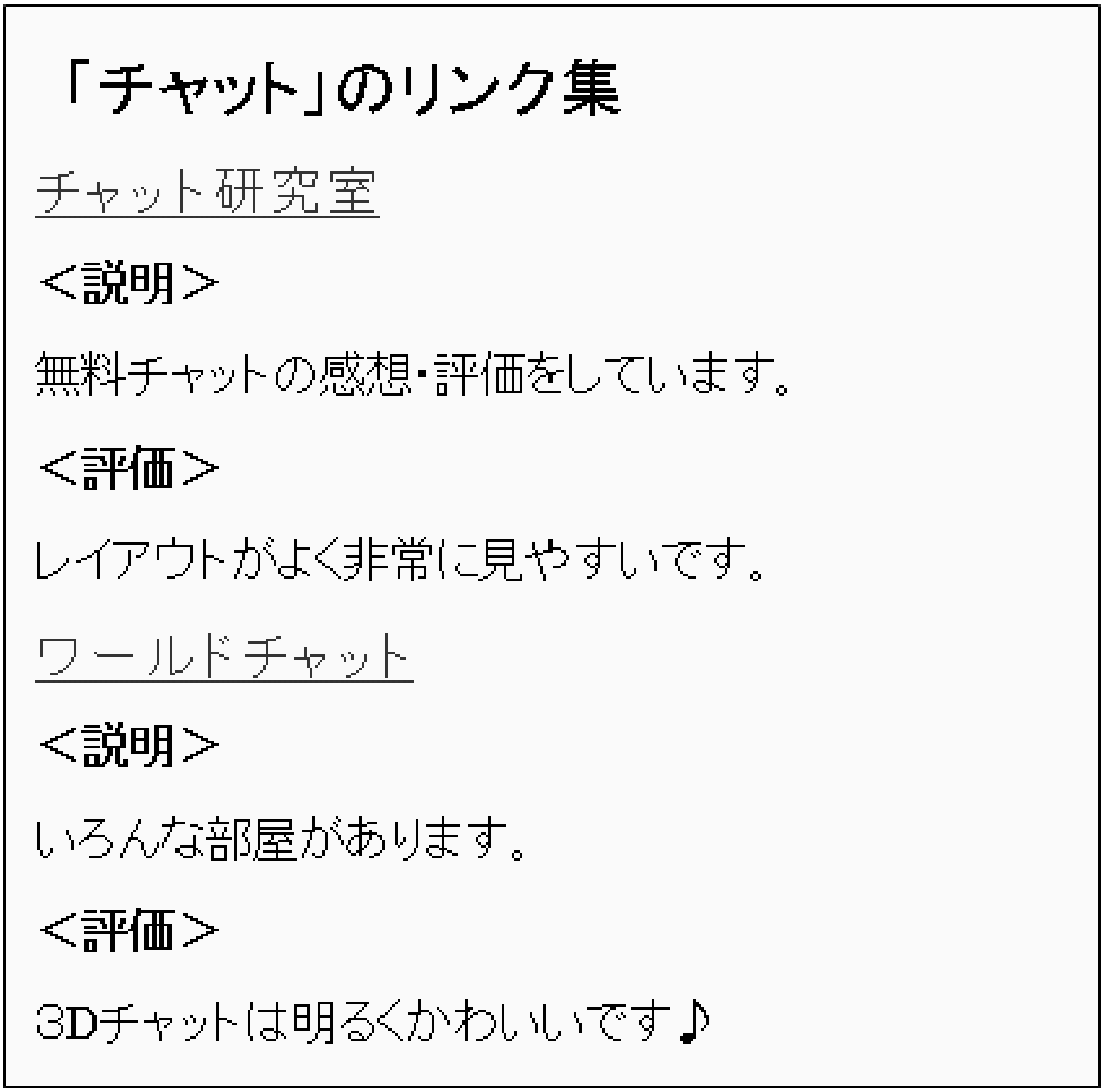

図 1.1: 出力する関連リンク集の例

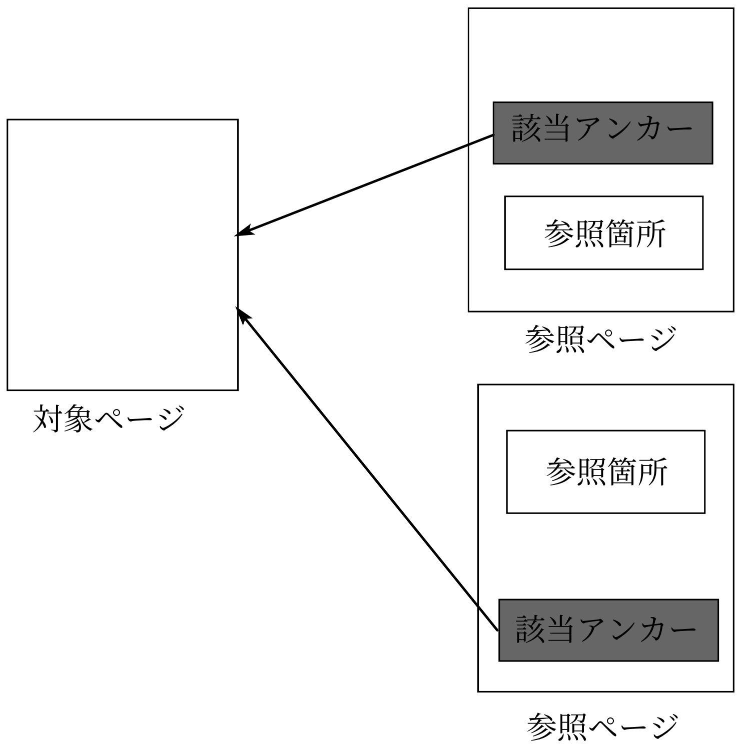

図 1.2: 対象ページと参照ページの定義

第2章　関連研究

2.1　Web ディレクトリの自動生成

佐藤らは水族館，動物園，博物館のような特定したカテゴリのための Web ディレクトリを生成する方法を提案している [1]．このシステムは，ユーザが入力したカテゴリワードからインスタンスを収集する．例えば，ユーザが水族館と入力すれば，Waikiki 水族館，オフォーツク水族館といったインスタンスを収集し，地方別に分類する．また Web ページから各インスタンスの名前，住所，電話番号，概要，郵便番号を抽出し提示する．このシステムは以下の 3 つのモジュールから成る．

- Name Collector（ユーザが入力したカテゴリワードに関する固有名を収集する.）
- Contents Editor（インスタンスのダイジェストページを作成する.）
- Organizer（同一性のチェック）

Name Collector は水族館の名前といった固有名を収集する．例えば，カテゴリワードの aquarium から，システムは aquarium の名前 (例：Waikiki Aquarium, Sea Life Park Hawaii,Monterey Bay Aquarium) を収集する．収集する方法として，カテゴリワードを主辞とする固有名を収集する．また，ある Web ページに水族館のリストがあり，そのリスト中にカテゴリワードが含まれていない固有名が存在した場合，それらもカテゴリワードに関する固有名と判断し，収集する．

Contents Editor はインスタンスのダイジェストページを作成する。また，ダイジェストページには名前，住所，電話番号，概要，郵便番号が提示される．

Organizer は収集したインスタンスの同一性をチェックし、目次を作成する。Contents Editor で獲得してきた名前，住所，郵便番号，電話番号を利用し，同一性のチェックを行う．例えば，2 つのインスタンスの名前が違っていても，住所が同じであるなら同一と判断する．もし，二つのインスタンスが同一であると判断した場合，2 つのインスタンスを併合する．また，郵便番号により地方別に分類し，階層的に整理した状態で出力する．

佐藤らは，ディレクトリ生成を自動的に生成し，Web 上から名前，住所，電話番号等を抽出している．本研究も同様に Web 上から情報抽出を行っている．しかし，本研究はディレクトリの自動生成ではなく関連リンク集の自動生成を目的としている．また，水族館や博物館の名前，住所ではなく，Web ページに関する記述を抽出する．

2.2　Webページの情報抽出

板橋らは，対象ページ自身から Web ページの情報を取り出すのではなく参照ページから取り出す事により，対象ページから取り出せない第三者の意見，評価等が取り出せると考え，参照ページから対象ページの記述を取り出す手法を提案している [2].

板橋らの手法は，最初に対象ページを指すアンカーを持つ参照ページを収集する．次に対象ページに関する説明，評価といった記述を参照ページから抽出し，ユーザに提示する．参照箇所の抽出する際には HTML タグを手掛かりとしている．具体的には，リストタグ，br タグ，テーブルタグ等を手掛かりとしている．リストタグの場合，アンカーの直前の li タグから，次の li タグまでを参照箇所として抽出している．br タグの場合，アンカー＋文字列＋ br タグというパターンが 3 回以上並んだとき．アンカーの次の文字列を参照箇所として抽出する．テーブルタグの場合，アンカーのセルの右側に参照箇所がある場合とアンカーが存在するセルの下に参照箇所がある場合がある．この場合，テーブルのレイアウトを判別し，セルの右または下のどちらに参照箇所があるかを判定し，抽出する．リストタグ，br タグ，テーブルタグ以外の場合，アンカーのすぐ前の HTML タグを先頭，アンカーのすぐ後の HTML タグを末尾と考えて参照箇所を取り出す．このとき，イメージタグ，文字修飾タグ，a タグ，コメントは無視する．また，br タグについては，最初に出現したときは無視し，2 回目の br タグを末尾とする．

また，Amitay らは，参照ページのアンカーの周辺に，参照先 Web ページの概要を紹介する文章が多く見られる事に着目し，そうした紹介文を要約文として利用する方法を提案している [3].

難波らは，データベースから関連する論文を自動的に収集し，人間が特定分野のサーベイ論文作成する作業を支援するシステムを提案している [4]．具体的には，ユーザーのキーワードにあった論文を検索し，検索結果はリスト表示される．このリスト中に参照・被参照関係の論文がある場合，論文間の参照・被参照関係のグラフを表示することができる．このグラフを辿ることで，論文間の参照・被参照関係を用いた検索が可能になる．

本研究では，板橋らの研究と同様に HTML タグを手掛かりとし，アンカー周辺を参照箇所として収集する手法をとる．しかし，板橋らの研究の手法だけでは，アンカーから離れている Web ページの情報を抽出することはできない．そこで本研究では，サイト名を用いて，アンカー周辺以外の情報を抽出する事を試みる．また，板橋らの研究は情報を抽出するだけである．本研究では取り出した情報をタイプ別に分類することを試みている．

2.3　評価文の抽出

2.3.1　評価表現の収集

小林らはテキストマイニングの技術を応用し，評価対象表現，属性表現，評価表現の共起パターンを利用して，これらの表現を効率的に収集することを試みている [5]．また，

共起パターンに基づく属性・評価表現の半自動的収集方法を提案している．例えば,「商品１の液晶は本当にきれい」という文の場合，評価対象表現は＜商品１＞であり，属性表現は＜液晶＞，評価表現は＜きれい＞となる．

表現の収集方法に関して具体的に説明する．まず，共起パターンと対象名・属性表現・評価表現の初期の辞書を用意する．属性表現・評価表現の辞書は，それぞれ正例辞書と負例辞書からなる．与えられたWeb文書に，共起パターンにマッチする部分が存在し，かつその表現のいずれかが辞書に存在するならば，他の部分を抽出する．初期の辞書の評価表現から，共起パターンを用いて新しい属性表現を収集でき，また逆に，属性表現から新しい評価表現を収集することができると考えている．このサイクルを繰り返す事によって，初期の辞書から大規模な辞書を構築する事が可能である．

収集方法の例を挙げる．

[属性] が・は・も・に・を [評価表現]

例えば，上記のような共起パターンと評価表現を用いて属性表現を収集する．まず，評価表現の正例辞書には「きれい」や「良い」といった表現が既に含まれているとする．この時，与えられたWeb文書中に下記のような文章があれば，そこから「液晶」や「デザイン」を属性表現の候補として獲得する事ができる．

商品１って液晶がとっても綺麗 !!
デザインも良く，気に入っています

また，長江らは，企業の製品コンセプトとユーザの製品評価文章とを比較するシステムについて提案している [6]．彼らは，Web上の製品評価サイトから製品評価文章を収集し，これらを肯定的な意見か否定的な意見かを判定するための評価語辞書を作成した．また，製品カタログに含まれる企業の製品の宣伝文句等を分析している．

村野らは，まず掲示板中の多くの記事の中から主観的評価が記述されている記事を選別する [7]．さらに評価対象が製品の特定の項目であるか，製品全体の総合的なものであるか，また別の製品と比較した評価であるか，比較対象を明示せずに肯定，否定を述べた評価であるかで主観的評価を分類し，整理してユーザに提示するシステムを提案している．

小林らは，文章の内容が「望ましい事態 (p)」なのか,「望ましくない事態 (n)」なのか，あるいはどちらでもないのか想定するために，ブーストストラップ的手法で，p/n辞書を獲得を試みている [8].

Riloff らは，主観的表現を取り出すことを目的として，ブートストラップ的手法で抽出パターンを学習している [9]．また，分類器によってタグなしテキストにラベルを貼ることにより，膨大な訓練データが自動的に生成される．この分類器は主観性分類器と客観性で分類器の２つあり，それぞれのタグなしテキストに「主観的」「客観的」といったラベルを貼る．ブートストラップ的手法で学習されたパターンが増えるにつれて，高い精度が維持され，再現率も向上すると考えている．

本研究では，小林らのように「望ましい事態」なのか「望ましくない事態」なのか区別するのではなく，ページの説明なのか評価なのか判別する．また，Riloff らや小林らの

ようにブートストラップ的手法を用いず，人手で評価文を判定するパターンを作成する．長江らと村野らの研究は，製品を対象とした，評価文または評価表現を抽出・分類することを目的としている．これに対し，本研究では，Web ページに関する記述を対象としている．すなわち，ある文が Web ページに関する説明なのか，あるいは第三者による Web ページの評価であるかを自動的に分類することを目指す．先行研究で得られた知見を参考に，このような判定を行うパターンを人手で作成することを試みる．

第3章　関連リンク集の自動生成

3.1　システム概要

本研究は，与えられたキーワードに関連するリンク集を自動的に作成する．本システムの概要を以下に述べる．

1. ユーザによるキーワード入力
2. キーワードに関するページの URL を収集
3. 収集したページの参照ページを収集
4. 参照ページから対象ページに関する記述を自動抽出
5. 得られた記述を説明文・評価文などに分類
6. 関連リンク集として出力

本研究では関連リンク集に提示するページを対象ページ，対象ページのリンクを貼っているページを参照ページ，参照ページ内にある対象ページに関する情報が記述された部分を参照箇所と定義する．

3.2　Web ページの収集

ここでは，3.1 節のステップ 1-3 を行う．

まず，ユーザーがキーワードを入力する．キーワードを Goo にかけ，キーワードに関する対象ページの URL を収集する．次に，各々の対象ページの参照ページを収集する．方法は Goo のリンク先 URL 検索を用いて行う．図 3.1 に例を示す．まず最初にユーザが「グルメ」と入力したとする．すると,「グルメ」に関連する Web ページ「グルメぴあ，ぐるなび，ぐるめピタ等」の URL を収集する．次に「グルメぴあ」,「ぐるなび」,「ぐるめピタ」の URL を Goo にかけ，リンク先 URL 検索を行い，それぞれの参照ページを収集する．

検索エンジンによって収集されたページの中には，リンク集に掲載するべきでないページも存在する．関連リンク集の自動生成において，リンク集に掲載するページの選別は重要な手続きである．しかし，本研究は，ページに関する説明の収集・分類に焦点をあてているため，Web ページの選別は行わない．

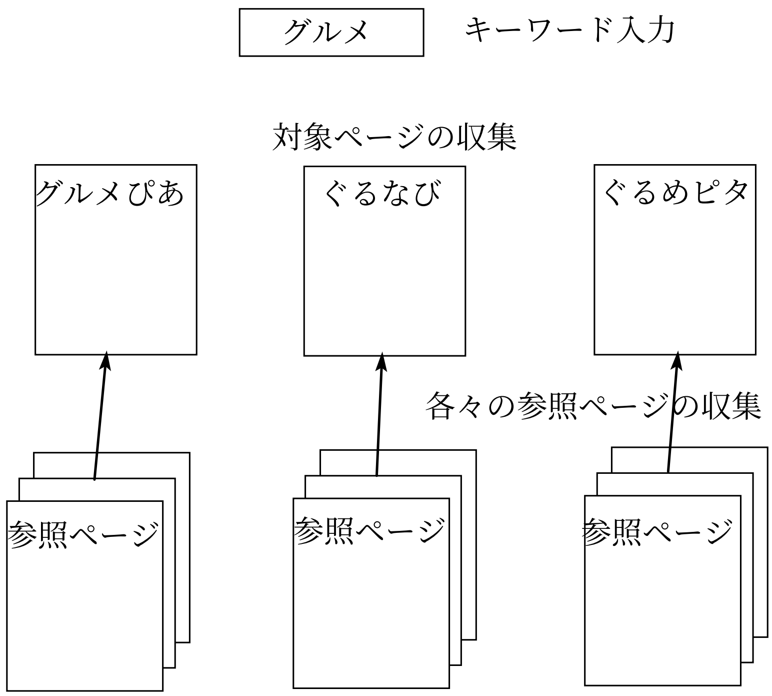

図 3.1: Web ページの収集

3.3 Web ページに関する情報の抽出

ここでは，3.1 節のステップ 4 を行う．

3.3.1 HTML タグを用いた方法

本項で述べる手法は板橋の手法 [2] と同じである．ここでは，その手法を簡単に紹介する．詳しくは文献 [2] を参照されたい．

まず，HTML タグを用いて，参照ページから参照箇所を抽出するアルゴリズムを図 3.2 に示す．初めに該当アンカーがナビゲーション目的での参照であるかどうかを判定する．ナビゲーション目的とは，主にサイト内リンクを示す．例えば，アンカー文字列が「戻る」とか「TOP へ」とかであるなら，その該当アンカーの周辺には対象ページに関する記述がないと考える．そこで該当アンカーがナビゲーション目的であるなら，参照箇所は存在しないと判定する．ナビゲーション目的の参照だと考えられる文字列は以下のリストに示す．

- 「トップへ」で終わる文字列
- 「ホーム（へ）」で終わる文字列
- 「ホームページへ」で終わる文字列

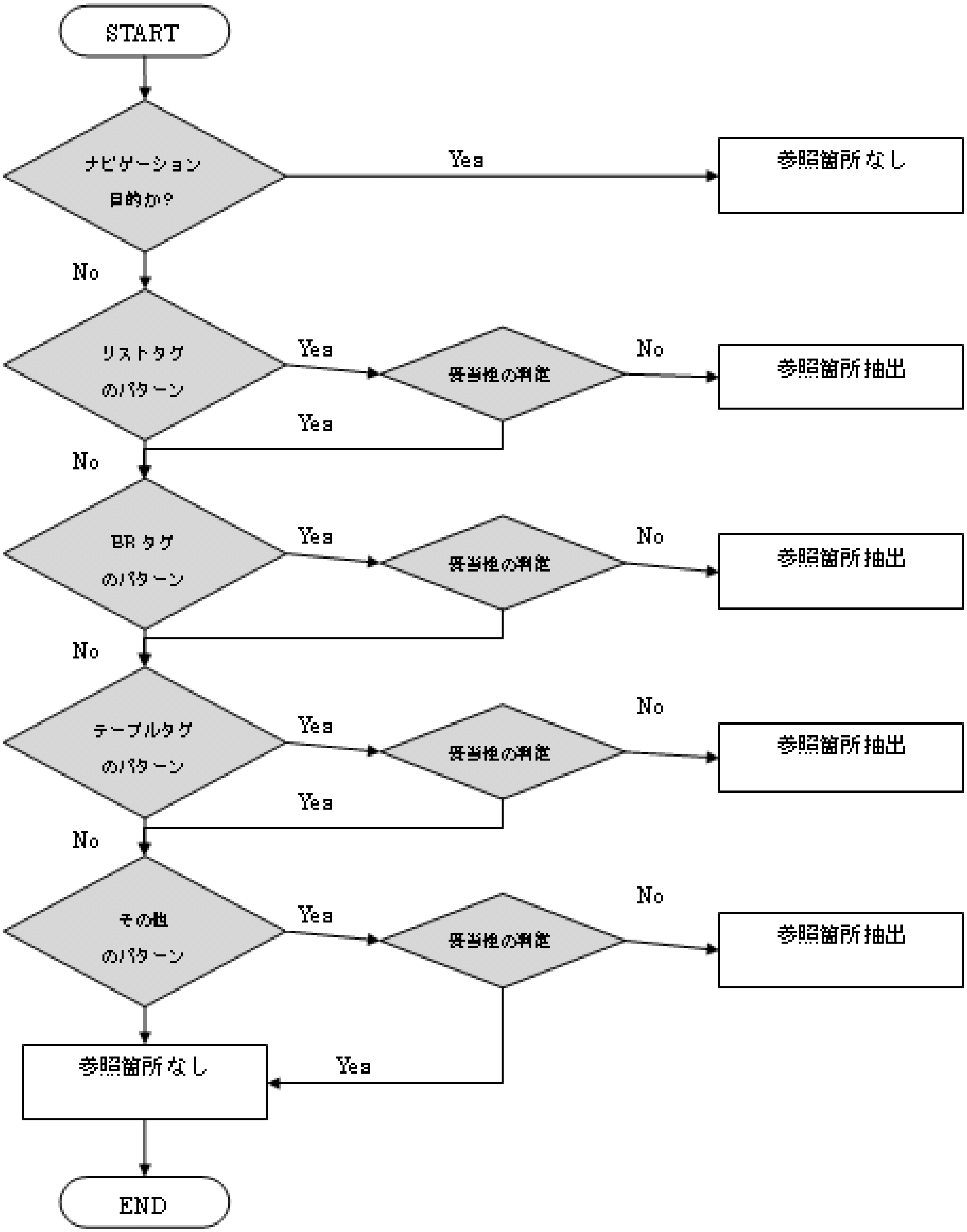

図 3.2: 参照箇所抽出アルゴリズム

- 「ホームページ」
- 「戻る」で終わる文字列
- 「もどる」で終わる文字列
- 「TOP（へ)」で終わる文字列
- 「back」
- 「home（へ)」で終わる文字列
- 「home page へ」で終わる文字列
- 「home page」

次に参照箇所の抽出を行う．抽出の際，リストタグ，テーブルタグ，br タグ，その他を手掛かりとする．

リストタグの場合

リストタグを手掛かりとして抽出する場合を述べる．ここで，下線が入っている部分を対象ページの該当アンカーとする．

> $< li >$ ぐるなび： 目的、地域、形態でお店を検索できます。関東、関西、全国版などがあります。
> $< li >$Yahoo!グルメ：地域、ジャンル、予算、キーワードから、メニューや平均予算などお店の情報が検索できます。
> $< li >$Walker グルメ：東京、横浜、関西、神戸、東海等のグルメ店が検索できます

上記のリストのように，該当アンカーの直前に li タグがある場合，その li タグから次の li タグまでの部分，すなわち「Yahoo!グルメ：地域、ジャンル、予算、キーワードから、メニューや平均予算などお店の情報が検索できます。」が参照箇所として取り出される．同様に，下記のように dl タグで書かれている場合が，該当アンカーの直前に dd タグがあるなら，それに対応する dt タグから dd タグまでの間を参照箇所として取り出す.

> $< dt >$ 住所（の一部）を入力して地図を検索できます。
> $< dd >$MapFanWeb

つまり「住所（の一部）を入力して地図を検索できます。」が参照箇所として取り出される．また，下記のように該当アンカーの直前に dt タグがある場合がある．

$<dt>$MapFanWeb
$<dd>$住所（の一部）を入力して地図を検索できます。

このときはアンカーの次にある dd タグ以降の文章を参照箇所として抽出する．

br タグの場合

次に br タグを手掛かりとした場合を述べる．下記のリストのようにリンクとリンク先ページの紹介が br タグで列挙されている場合，アンカーと次の br タグの間にある文章「パソコンと周辺機器の価格が一覧できます。パソコンのメーカーやスペック等から最安値などが検索できます。」を取り出す．

楽天：パソコンと家電、AV 機器などのスペック比較ができ、各ショップの価格一覧が検索できます。パソコン・家電の売れ筋ランキングもあります。$
$
価格.com：パソコンと周辺機器の価格が一覧できます。パソコンのメーカーやスペック等から最安値などが検索できます。 $
$
BestGate：パソコンや周辺機器の再安価格や在庫、２週間の値動きなどが見られます。 $
$

また，br タグを使わずに他のページへのリンクを列挙する場合もある．該当アンカーの前後に，該当アンカーを含めて他のページへのリンクが 3 回以上続く場合，該当アンカーとその直後の文字列を参照箇所として取り出す．例として以下のリストの場合を述べる．

アンカー$_1$　文字列$_1$　$
_1$
アンカー$_2$　文字列$_2$　$
_3$
アンカー$_3$　文字列$_3$　$
_3$

ここで，文字列$_i$ と $
_i$ は空でも良い．マッチすれば該当アンカーとその直後の $
$ タグもしくは他のページのアンカーまでの部分を参照箇所として取り出す．例として アンカー$_2$ が該当アンカーだとすれば,「アンカー$_2$　文字列$_2$」が参照箇所として取り出される．

テーブルタグの場合

次にテーブルタグを手掛かりとした場合を述べる．例えば下記のようにアンカーと参照箇所が同じ行にある場合，アンカーを含むセルの右側のセルにアンカー以外の文字列が記述されていれば，その文字列を参照箇所として取り出す．
次に以下のように，アンカーと参照箇所が交互にある場合，テーブルの同じ列にアン

MapFanWeb	住所（の一部）を入力して地図を検索できます。
MapBlast!	世界の地図が検索できます。
MAPION	日本全国の地図が目標物で検索できます。

ぴあ
チケットの前売り情報などがあります。
歌ネット
曲名・歌手名・作詞者名と 歌い出しの詞から歌詞が検索できます。

カーとアンカー以外の文字列が交互に並んでいた場合，該当アンカーの下のセルにある文字列を参照箇所として抽出する．

　また，数少ないパターンだが，以下のように，アンカーが横に並んでいるタイプも見られた．そこで，該当アンカーと同じ行にあるセルが，7 割以上がアンカーであるなら，該当アンカーの下のセルに記述された文字列を参照箇所として抽出する．

アンカー$_1$	アンカー$_2$	アンカー$_3$
参照箇所$_1$	参照箇所$_2$	参照箇所$_3$

その他の場合

　今まで述べたリストタグ，br タグ，テーブルタグのパターンにどれも当てはまらない場合，該当アンカーの近辺を参照箇所として抽出する．参照箇所の先頭と末尾は HTML タグで決定する．つまり，該当アンカーの前に存在する HTML タグを先頭にし，該当アンカーのすぐ後の HTML タグを末尾とする．この時，文字修飾タグや以下のタグは，参照箇所の境界を決める際に無視する．

- 文字修飾タグ（font,b,i 等）
- image タグ
- a タグ
- コメント
- br タグ（ただし，2 回目の br タグは無視しない）

例を以下に示すと，

IT media< *br* >
インターネットや新製品情報などのニュースなどがあります。

この場合, テーブルの一つのセルの中に該当アンカーが存在する．よって，該当アンカーの前後にあるテーブルタグを検出した時点で参照箇所の境界を決められる．

次に, 抽出した参照箇所の妥当性を判定する．つまり，自動的に抽出された参照箇所が対象ページに関する情報として妥当であるかどうか判定する．抽出された参照箇所が以下の条件を一つでも満たす場合，参照箇所として抽出しない．

- 英字，記号（○×ー☆等）からなる 5 文字以下の文字列
- 参照箇所が対象ページの URL の部分文字列である（goo.ne.jp）

3.3.2　サイト名を手掛かりとした参照箇所の抽出

前項では HTML タグを利用し，アンカー近辺に存在する参照箇所を収集する方法を述べた．しかし，アンカー近辺以外で対象ページに関する情報が記述されているページが存在する．前項の方法ではこのようなアンカーから離れた対象ページに関する情報を取り出すことはできない．そこで，アンカー近辺以外の Web ページに関する情報の抽出を試みる．

アンカー周辺以外で書かれている Web ページに関する情報は，対象ページのサイト名を省略せずに書いてあるものが多く見られた．すなわち，対象ページのサイト名を手掛かりとしてアンカー近辺から離れた参照箇所を取り出せるのではないかと考えた．

まず，対象ページのサイト名を特定する．対象ページのアンカー，つまり該当アンカーのアンカー文字列には，サイト名が書かれていることが多い．そこで，サイト名として，該当アンカーのアンカー文字列を抽出する．ただし，あまりにも長い文字列はサイト名でない事がほとんどなので，サイト名として抽出しない．ここでは，文字列の長さが 13 文字以上の場合はサイト名でないと判断し，抽出しない．また，サイト名を表す文字列は一つとは限らないので，複数のアンカーからサイト名を複数抽出する．

次に，サイト名を手掛かりとして参照箇所の抽出を行う．具体的には，サイト名を表す文字列を含む文を抽出する．参照箇所の境界は，HTML タグを手掛かりとし決定する．以下に例を示す．

$< p >$ サイト A はかなり充実したコミュニティサイト。人もたくさんあつまっている。オススメできる。サイト B も有名であるが，サイト A の方が安定している．
$< br >$

いろいろな記述や表や図等

サイト A
サイト B

この例では，アンカーの近辺に参照箇所がなく，離れた場所に Web ページの記述が書かれている．まず，サイト名「サイト A」の文字列の前にある HTML タグを探し，参照箇所の先頭とする．同様に,「サイト A」の文字列の後に存在する HTML タグを探し，参照箇所の末尾とする．つまり，この例では，$< p >$ から $< br >$ の間にある「サイト A はかなり充実したコミュニティサイト。人もたくさんあつまっている。オススメできる。サイト B も有名であるが，サイト A の方が安定している.」が参照箇所として抽出される．

3.4　参照箇所の分類

本節では 3.1 節のステップ 5 について述べる．

本研究では，抽出してきた参照箇所を無秩序に列挙するのではなく，情報をタイプ別に分類・整理し，ユーザに提示する事によって，どの Web ページがユーザにとって有用か把握しやすくなると考えた．本節では，参照箇所のカテゴリとその自動分類手法について述べる．

3.4.1　参照箇所のカテゴリ

Web ページに関する情報には，Web ページの内容を説明する文と Web ページに対する評価や意見を述べる文があった．また，Web ページの内容を説明する文の中で最も多かったのは，Web ページに関する一般的な説明であった．また，どんな機能があるのかを説明する文も多数見られた．一方，評価や意見を表わす文の中では，Web ページの利便性や使いやすさ等を表す文と，Web ページの規模または情報量等を示す文が多数見られた．そこで，本研究では，参照箇所を以下の 5 つのカテゴリに分類する．

評価：利便

Web ページの利便性，使い勝手，見やすさといった記述が参照箇所中にある場合,「評価：利便」のラベルを付与する．

例：製品価格を調べるときに便利だ

評価：情報量

Web ページの規模，情報量といった記述が参照箇所中にある場合,「評価：情報量」のラベルを付与する．

例：豊富な情報量を持つ

評価：その他

Web ページの利便性や情報量といった記述以外の評価が参照箇所中にある場合,「評価：その他」のラベルを付与する．例えば，Web ページのデザインや安定性，速度に対する評価等はこのラベルが付与される．

例：楽しくチャットができます

説明：機能

その Web ページで可能な事や，どういった機能があるか，または機能の内容といった記述が参照箇所中にある場合,「説明：機能」のラベルを付与する．

例：地域別検索ができる

説明：記述

このカテゴリは Web ページに関する一般的な説明を意味する．

例：〇〇株式会社が運営するサイト

3.4.2 参照箇所の分類

参照箇所を上記５つのカテゴリに分類する．各カテゴリの特徴的な表現をつかむことで参照箇所をカテゴリ別に分類することを試みる．具体的には，５つのカテゴリ別にパターンを用意し，パターンマッチに成功すればそのカテゴリのラベルを付与する．分類のためのパターンを以下に示す.

カテゴリ 評価：利便のパターン

[評価:利便] に属する文を判別するパターンは以下の通りである．

- [評価表現 (利便)]

  [評価表現 (利便)] とはページの利便性を示すキーワードを表わす．その一覧表を表3.1 に示す．利便性，使いやすさに関する評価文の多くは,「便利」「使いやすい」といった利便性を示す評価表現が含まれているだけで，Web ページの利便性を評価している文と判定できるものが多かった．そこで,「便利，使いやすい」といった [評価表現 (利便)] に属する単語を含むなら，[評価：利便] カテゴリのラベルを付与する．このパターンで分類できた文の例を示す．以下の文章は,「便利」といった利便性の評価表現が含まれているので，この文章に [評価：利便] のラベルが付与される．

  ex．製品価格を調べるときに便利だ

表 3.1: 評価表現 (利便)

評価表現 (利便)
便利, べんり, 便利だ, べんりだ, 使いやすい, 見やすい, 楽, 楽だ, 役立つ, 役に立つ, 楽チン,

カテゴリ 評価:情報量のパターン

[評価:情報量] に属する文を判別するパターンは以下の通りである．

- [属性表現 (情報量)](が | は | も | に | を | など)− > [評価表現 (情報量)]

  [属性表現 (情報量)] とは,「コンテンツ」「利用者数」など，Web ページに含まれる情報を表わす名詞である．一方，[評価表現 (情報量)] は，それらの量を表わすような表現である．[属性表現 (情報量)]，[評価表現 (情報量)] の定義をそれぞれ表 3.2，表3.3 に示す．また，記号「− >」は属性表現と評価表現との間に係り受け関係が成立することを表わす．属性表現と評価表現が隣接していなくてもよい．例えば，以下の例文はこのパターンにマッチするが,「コンテンツが」と「多い」の間に係り受け関係が成立するなら，その間に別の単語があってもよい．また，[属性表現 (情報量)] はこの例文では「コンテンツ」であり，[評価表現 (情報量)] は「多い」である．

  ex. コンテンツが多い

- [属性表現 (情報量)](が | は | も | に | を | など) + [評価表現 (情報量)]$_{\text{連用形}}$

  [評価表現 (情報量)]$_{\text{連用形}}$ とは情報量の評価表現の連用形のみを示す．上記のパターンの記号「+」は「− >」とは違い，属性表現と評価表現が隣接していなければならないことを表わす．以下の文「ジャンルが豊富で便利だ」において,「ジャンル」

表 3.2: 情報量の属性表現

[属性表現 (情報量)]
検索手段，情報，情報量，コンテンツ，ジャンル，利用者，利用者数，旅行社，旅行社数，会員，会員数，掲載，掲載数，宿泊情報，ページ，ホームページ

表 3.3: 評価表現 (情報量)

評価表現 (情報量)
大きい, 小さい, 多い, 少ない, 膨大, 膨大だ, 満載, 充実, 豊富, 豊富だ, たくさん, たくさんだ, 沢山, 沢山だ, 最大級, 最大, さまざま, さまざまだ, 様々, 様々だ, 盛りだくさん, 盛りだくさんだ, 幅広い, 十分, 十分だ, たくさん+ある, (数字)

は [属性表現 (情報量)]（表 3.2）であり,「豊富」は [評価表現 (情報量)]（表 3.3）であるので，この例文は上記のパターンにマッチする．ちなみに,「便利」は [評価表現 (利便)]（表 3.1）であるので，この文には [評価:情報量] の他に [評価:利便] のカテゴリも付与される．

ex. ジャンルが豊富で便利だ

- [評価表現 (情報量)][属性表現 (情報量)]

  「充実したコンテンツ」や「膨大な情報」というように Web ページの情報量を評価する文が見受けられたので，上記のパターンを用意して，判別した．以下の例文,「豊富」は [評価表現 (情報量)] であり,「情報量」は [属性表現 (情報量)] であるので, 上記のパターンにマッチする．

  ex. 豊富な情報量を持つ

- [属性表現 (情報量)](が | は | も | に | を | など) + 数量表現

  数量表現とは数字を表す．以下の例文の「掲載数」は [属性表現 (情報量)] であり,「1000」は数量表現であるので，上記のパターンにマッチし，[評価:情報量] のカテゴリが付与される．

  ex. 掲載数が 1000 件ある。

- 満載 + (< 文末 > |。|. |? | !)
  満載だ
  満載の + [ページ表現]

  参照箇所中に「ー情報満載」といった文が多く見られた．そこでこのような文を判別するために上記のパターンを用意した.「ページ表現」とは「ーページ，ー HP」のようにWebページを表わす表現を指す.

表 3.4: ページ表現

ページ表現
ページ, ホームページ, サイト, ＨＰ, リンク集, ＬＩＮＫ集

- 数量表現 + [情報量接尾語]

  [情報量接尾語] とは「ー件」,「ー店」のように，主にWebページに掲載されているコンテンツの数量表現の後ろにつく単位を示す．情報量を含む評価文のうち，このパタンにマッチする文は多く見られた．[情報量接尾語] のリストを表 3.5 に示す．以下の例文はこのパターンにマッチする文である.「500 店」は「数量表現 + [情報量接尾語]」の組み合わせであるので，以下の例文に [評価:情報量] のカテゴリが付与される．

  ex. レストランが５００店登録されている

表 3.5: 情報量を表わす接尾語

情報量接尾語
件, 都市, カ所, 店, 人, 施設

カテゴリ 評価:その他のパターン

- [名詞](が | は | も | に | を)− > [評価表現 (その他)]

  「評価表現 (その他)」とは，[評価表現 (利便)] と [評価表現 (情報量)] 以外の形容詞，または表 3.6 のいずれかの単語を指す．このパターンにマッチする例文を以下に示

す.「機能」は名詞であり,「魅力的」は [評価表現 (その他)] であるので, [評価:その他] のカテゴリが付与される. ちなみに,「使いやすい」は [評価表現 (利便)] であるので, [評価:利便] のカテゴリも同時に付与される.

ex. 使いやすい検索機能が魅力的だ

表 3.6: 評価表現 (その他)

評価表現 (その他)
評判, 評判だ, 優れる, 魅力, 魅力的, 魅力的だ, 多様, 多様だ, 最適, 最適だ, 驚く

- [評価表現 (その他)] + [ページ表現]

  「透明感があるデザインの素晴らしいページ」,「風景写真が美しいサイト」,「素朴感ありのかわいい HP」のように Web ページのデザインを評価した文や,「面白いページ」といった評価文がまれに見られた. そこで上記のパターンを用意した. このパターンにマッチする例文を以下に示す.「素晴らしい」は [評価表現 (その他)] であり,「ページ」は [ページ表現] であるので, [評価:その他] のラベルが付与される.

  ex. 素晴らしいページだ。

- ので + (おすすめ | お薦め | お奨め | お勧め)
  オススメ

  「美味しくてお洒落なお店に詳しい紹介されているのでお勧めです.」や「このサイト有料なのでお薦めしません.」といったように, 記述者が Web ページを推薦する文があった. そこで上記のパターンを用意し, 記述者の推薦文に [評価:その他] のラベルを付与する.

- (もう少し | もっと) + [評価表現 (その他)] + [名詞] + (が欲しい | のほうがいい)

  以下の例文のように第三者の Web ページに対する要求や希望, 改善点といった意見があった. そこで上記のパターンを作成し, 要求, 希望, 改善点といった意見文を [評価:その他] として判別した.

  ex. もっとかわいらしいデザインのほうがいい

- (楽しく | 早く) + [動詞]

  「楽しく遊べます」や「情報をどこよりも早く入手できます」といった文があった場合，上記のパターンを用いて [評価:その他] に判別する．このパターンにマッチする例文を以下に示す.

  > ex. 楽しくチャットができます。

カテゴリ 説明:機能のパターン

- [属性表現 (機能)](が | は | も | に | を)– > [可能表現]

  「可能表現」とは「ーできる，ーくれる，ー可能，ーれる，ーられる」といった表現である．[属性表現 (機能)] とは,「検索」「予約」など，Web ページに関する機能を表わす名詞である．[属性表現 (機能)] の単語リストを表 3.7 に示す．例えば，以下の例文では,「印刷」は [属性表現 (機能)] であり,「ーできる」は [可能表現] であるので，このパターンにマッチし，この文に [説明:機能] のカテゴリが付与される．

  > ex. ルートマップの印刷ができる

表 3.7: 機能の属性表現

[属性表現 (機能)]
表示, 検索, 登録, ダウンロード, 予約, 確認, 利用, 設定, 手続き, 手続, 番号, プリント, 印刷, 道順, ルート, 変更

- (で | まで)– > [可能表現]

  可能を表わす表現に「ーで」や「ーまで」が係る文に，Web ページの機能を説明するものが多く見られたため，このパターンを設けた．上記のパターンにマッチする文を以下に示す.

  > ex. ワンクリックで予約できる

- [属性表現 (機能)] + (できる | 可能)

  「ー検索できる」「ー予約できる」といった表現で，Web ページで可能なことを記述されている場合があったので，上記のパターンを作成した．以下の例文の「検索」は [属性表現 (機能)] であり,「[属性表現 (機能)] ＋できる」といった型をしているので，上記のパターンとマッチし，[説明:機能] のラベルを付与する．

  > ex. 検索できる．

- [属性表現 (機能)] + (| を | も) + することが + (できる | 可能)

  「ー予約をすることができる」「ー設定をすることができる」といった表現で，Web ページで可能なことを記述されている場合があったので，上記のパターンを作成した．以下の例文も「[属性表現 (機能)] ＋を＋することが+できる」といった型をしているので，上記のパターンとマッチし，[説明:機能] のラベルを付与する．

  ex. 検索をすることができる

- [機能表現語]

  機能を表わす単語の中で，その単語が含まれているだけで [説明:機能] であるといえる場合が多数見受けられた．そこで，このような単語を [機能表現語] と定義し，単語リストを作成した．[機能表現語] のリストを表 3.8 に示す．

表 3.8: 機能表現語

[機能表現語]
メールサービス, メール配信サービス, 予約+システム, 応募+システム, サービス+開始

- (サービス | サービス提供) + を + 開始

  IT 関連の記事等で，Web ページのコンテンツサービスの開始に関するアナウンスが書かれていることがある．このような文も「説明:機能」としてラベルを付与する．

- (オンライン $|Web|$ インターネット | ネット) + で + (提供 | 開始 + 〜できる)

  「カラオケをインターネットで提供」や「目的地まで行く際の経路が検索できる機能をオンラインで提供」といった表現で Web ページの機能を説明する文があった．そこで，上記のパターンを用意し，このような説明文に対して [説明:機能] のカテゴリを付与できるようにした．

カテゴリ 説明:記述のパターン

上記の４つのカテゴリが一つも割り当てられなかった場合，[説明：記述] が付与される．以下にこのカテゴリに属する例文を示す．この文は今まで述べてきた 4 つのカテゴリのどのパターンにもマッチしない．したがって，この例文には [説明：記述] のカテゴリが付与される．

- オンラインソフトの紹介をしている

本研究では，3.4.1 項において，参照箇所に与えるべき５つのカテゴリを提案した．Web ページに関する記述によっては，これらのうち複数のラベルが付与される場合がある．例えば，どんな機能があって，またその機能の利便性を評価する文があった場合,「説明：機能」と「評価：利便」のラベルが付与される．

「属性表現」「評価表現」といった手掛かり語として有効なのは，一つのカテゴリと密接に関係のある単語である．例えば，[評価：利便] について考えると,「便利」といった単語は有効な手掛かりになると考えられる.「便利」という単語は，その文に出現するだけで利便性を示すと考えられるからである．[説明：機能] については,「機能」,「検索」といった単語は有効である.「検索」自体が，Web ページの機能を表わすからである．また，形容詞も評価文の手掛かりとなる事が多かった．

第4章　評価実験

本章では，提案手法の評価実験を行う．4.1 節では参照箇所抽出の評価実験，4.2 節では参照箇所の自動分類の評価実験について述べる．

4.1　参照箇所抽出実験

HTML タグを用いて参照箇所を抽出した場合と，サイト名を手掛かりした場合とで，どれだけ参照箇所を抽出できたかを比較した．サイト「BBgames」と「壁紙.com」を対象ページとして実験を行った．実験結果を表 4.1 に示す．表 4.1 から，参照箇所の 30〜40%がサイト名を手掛かりとした手法で抽出されている．このことから，2 つの対象ページについてのみしか調べていないが，HTML タグのみを用いる手法と比べて，サイト名も手掛かりとする本研究の方がより多くの参照箇所を抽出できるといえる．

表 4.1: 分類に関する実験結果

対象ページ	HTML タグ	サイト名
BBgames	24(69%)	11(30%)
壁紙.com	49(60%)	32(39%)

4.2　カテゴリ分類の実験準備

本研究の最終目的はリンク集の自動生成である．そのため，どのようなテーマのリンク集も自動生成しなければならない．そこで，ポータルサイトのテーマとして 17 テーマを設定し，それぞれについて 1，2 個，合計 21 個のサイトをリンク集に掲載するページとして選んだ．テーマと Web ページの一覧を表 4.2 に示す．これらのページを対象ページとし，提案手法によって 401 件の参照箇所を集めた．これらの 401 件の参照箇所は 3.4 節で述べたカテゴリ分類のためのパターンの作成に用いたものである．集めた参照箇所には手作業でカテゴリのラベルを付け，正解のカテゴリ付きの参照箇所リストを用意した．このデータを用いて参照箇所の自動分類のクローズドテストを行った．この際，文の形態素解析には JUMAN を用い，係り受け解析には KNP を用いた．

表 4.2: クローズドテストのテーマと対象ページ

テーマ	対象ページ
オンラインソフト	窓の杜 (http://www.forest.impress.co.jp/)
地図	MapFan Web(http://www.mapfan.com/)
メッセンジャー	MSN Messenger(http://messenger.msn.co.jp/)
天気予報	ワンクリック気象情報サイト (http://tenki.jp/)
映画	シネマスクランブル (http://cinesc.cplaza.ne.jp/)
旅行	旅行リンク (http://www.ryokolink.com/)
旅行	旅行の窓口 (http://www.mytrip.net/)
番組表	Yahoo!テレビ (http://tv.yahoo.co.jp/)
あるある大事典	発掘！あるある大事典 WEBSITE(http://www.ktv.co.jp/ARUARU/)
ニュース	Yahoo!ニュース (http://headlines.yahoo.co.jp)
無料ホームページ	楽天広場 (http://plaza.rakuten.co.jp/)
グルメ	ぐるなび (http://www.gnavi.co.jp)
グルメ	グルメぴあ (http://g.pia.co.jp/)
オークション	Yahoo!オークション (http://auctions.yahoo.co.jp/)
懸賞	ふくびき.com(http://www.fukubiki.com/)
懸賞	Chance It!(http://www.chance-it.com)
懸賞	MyID(http://www.myid.ne.jp/)
TV 局	フジテレビ (http://www.fujitv.co.jp/)
メールマガジン	まぐまぐ (http://www.mag2.com/)
HTML タグ	HTML タグであそぼう (http://www4.osk.3web.ne.jp/ kitayan/)
鑑定	ハニホー！(http://hanihoh.com/)

さらに，カテゴリ分類のパターンの作成に用いたデータとは異なる対象ページと参照箇所の組を用意し，同様に人手で正しいカテゴリを付与した．このデータを用いてカテゴリ分類の評価のためのオープンテストを行った．オープンテストでの対象ページ数は 16，参照箇所の数は 344 である．また，テーマ数は 11 である．オープンテストで使用したテーマとサイトを表 4.3 に示す．

表 4.3: オープンテストのテーマと対象ページ

テーマ	対象ページ
星野仙一	星野仙一のトラトラトラ (http://hoshino.ntciis.ne.jp/)
価格	価格.com(http://www.kakaku.com/)
コミュニティサイト	cafesta(http://http://www.cafesta.com/)
コミュニティサイト	この指とまれ (http://www.yubitoma.or.jp)
占い	動物占い (http://www.noracom.net)
育児	ベネッセウィメンズパーク (http://women.benesse.ne.jp/)
懸賞	ポイントメール (http://www.pointmail.com/)
懸賞	笑う懸賞生活 (http://www.warau.jp/)
懸賞	フルーツメール (http://www.fruitmail.net/)
懸賞	懸賞のつぼ (http://www.tubox.com/)
CM	CM サイト (http://www.cmsite.co.jp/)
ペット	ペット大好き！(http://www.petoffice.co.jp/)
お薬	みのりの広場 (http://www.eminori.com/)
お薬	おくすり 110 番 (http://www.jah.ne.jp/ kako/)
求人情報	ハローワークインターネットサービス (http://www.hellowork.go.jp/)
地図	Mapion(http://www.mapion.co.jp/)

クローズドテストの結果として，カテゴリ別ならびにテストデータと全体の精度と再現率を表 4.4 示す．また，実験データとして使われた参照箇所には，広告など，Web ページに関する記述でないものも含まれる．「全体 (ページ)」はそのようなページに関する記述でないものを除いたときの全体の精度と再現率である．表 4.5 にオープンテストの結果を示す．表に記載した項目は表 4.4 と同じである．

表 4.4: カテゴリ分類の実験結果 (クローズドテスト)

カテゴリ	精度	再現率
評価：利便	0.85 (23 / 27)	0.77 (23 / 30)
評価：情報量	0.88 (30 / 34)	0.77 (30 / 39)
評価：その他	0.31 (5 / 16)	0.42 (5 / 12)
説明：機能	0.64 (67 / 104)	0.76 (67 / 88)
説明：記述	0.77 (184 / 236)	0.73 (184 / 252)
全体	0.74 (309 / 417)	0.66 (309 / 467)
全体 (ページ)	0.84 (309 / 368)	0.73 (309 / 421)

表 4.5: カテゴリ分類の実験結果 (オープンテスト)

カテゴリ	精度	再現率
評価：利便	0.89 (8 / 9)	0.73 (8 / 11)
評価：情報量	0.67 (8 / 12)	0.53 (8 / 15)
評価：その他	0.75 (3 / 4)	0.27 (3 / 11)
説明：機能	0.68 (48 / 71)	0.71 (48 / 68)
説明：記述	0.71 (167 / 236)	0.78 (167 / 213)
全体	0.70 (234 / 332)	0.63 (234 / 372)
全体 (ページ)	0.84 (234 / 278)	0.74 (234 / 318)

4.3　考察

以下，各カテゴリ毎に結果を考察する．

評価：利便

クローズドテストと比べて，オープンテストの精度が 0.04 上がった．また，再現率が 0.04 下がった．データ数は少ないが，[評価表現 (利便)] も属する単語があるかないかという単純なパターンだけで判別ができ，このカテゴリを誤って付与することも少なかった．誤って付与した中には「Web ページ以外の記述」が含まれている場合がオープンテスト，クローズドテストとも見られた．

評価：情報量

クローズドテストより，精度が 0.21 下がり，再現率は 0.25 下がった．この大きな原因は手掛かり語の少なさである．[評価:情報量] の手掛かり語を増やし改善する必要性がある．

評価：その他

クローズドテストと比べて，オープンテストの精度が 0.44 上がった．また，再現率は 0.15 下がった．再現率は 0.27 で全カテゴリの中で最悪の結果となった．このカテゴリに関しては定義をより明確化する必要がある．例えば，[評価:その他] には Web ページのデザインの評価や Web ページの安定性，更新頻度について評価している記述があり，このような記述に対して独自のカテゴリを与えることを検討すべきである．

説明：機能

クローズドテストと比べて，オープンテストの精度が 0.04 上がった．また，再現率は 0.05 下がった．評価:情報量のカテゴリと同様に，手掛かり語を増やせば改善できる箇所が見られた．

説明：記述

クローズドテストに比べて，オープンテストの精度が 0.07 下がった．また，再現率は 0.05 上がったオープンテストとクローズテストの「全体」と「全体 (ページ)」の精度，再現率がほとんど変わっていない．しかし，[説明:記述] を除いた 4 つのカテゴリの再現率はいずれも下がっている．このことにもかかわらず,「全体 (ページ)」の再現率が変わらなかった原因として，[説明:記述] の再現率が向上したことが挙げられる．本システムでは，[評価:利便][評価:情報量][評価:その他][説明:機能] のどれにも属さない場合，[説明:記述] を

付与する．しかし，実際の Web ページに関する記述には，[説明:記述] と [評価:利便] といったようにカテゴリが複数付与される参照箇所がある．このとき，人手では [説明:記述][評価:利便] というように 2 つのカテゴリが与えられるが，システムは [評価:利便] を付与したとき，[説明:記述] を付与しない．クローズドテストでは，[説明:記述] と他の 4 つのカテゴリのいずれかを共に正解カテゴリとする参照箇所が多く見られた．このことが，[説明:記述] に対する再現率を低下させる大きな要因となっていた．一方，オープンテストでは，[説明:記述] と他のカテゴリを共に正解カテゴリとする参照箇所がほとんどなかった．この結果，みため上，オープンテストはクローズドテストと比べて [説明:記述] の再現率が向上した．これは，テストデータの数が少ないこと，テストデータの選び方に偏りがあったためと考えられる．したがって，評価を行うページの選別については検討を要する．

今回のカテゴリ分類の実験で用意したデータ量は十分ではない．今後，実験データを増やす必要がある．また，本研究では「説明：記述」に関してパターンと手掛かり語を用意しなかった．そのため,「『Exif NaviforMapFan』詳細やダウンロードはこちら」といったページに関する情報でない記述も「説明：記述」のラベルを付与することになった．よって,「説明：記述」にもパターンと手掛かり語を用意し,「説明：記述」に属する有用な情報と不要な情報を区別する必要がある．特に，宣伝を目的とした記述も度々抽出されていた．これは，第三者による意見ではない．よって，宣伝文句を判別するための手掛かり語やパターンを用意し，Web ページに関する記述でないことを判別する必要がある．

また，手掛かり語やパターンを大量に人手で用意するのは困難である．Riloff らの手法 [9] のようにブートストラップ的手法を用いて手掛かり語を収集する方法を検討する必要がある．

また，長い記述には有用な情報が書かれている事が多いが，提示される側としては非常に見にくい．よって，HTML タグを境界として抽出された形式的なセグメントではなく，もっと細かい単位で参照箇所を分解する必要がある．その際には，参照箇所中での語同士の意味的つながりを考える必要がある．

第5章　おわりに

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，Webページに関する情報の抽出し，それらを情報の種類に応じたカテゴリに分類することを試みた．Webページの情報を分類するために，手掛かり語とカテゴリの分類パターンを作成した．評価実験の結果,「評価：利便」,「説明：機能」のカテゴリの付与については7割以上の再現率を得ることができた．しかし，まだまだパターンと手掛かり語が足りない．

5.2　今後の課題

今回，作成したパターンと手掛かり語だけでは十分ではないので，追加する必要がある．また，HTMLタグとサイト名を手掛かりとし抽出された参照箇所には有用な情報が多く見られた．しかし，Webページに関する第三者の記述とはいえない，不要な情報も多数見られた．そのため，パターンや手掛かり語を用いて，抽出された参照箇所から，ユーザに提示すべきでない記述を取り除く必要がある．

次に，関連リンク集に掲載するWebページに関する記述は，参照ページから参照箇所を抽出する方法とった．しかし，Webページに関する評価文を記載している参照ページが少なかった．今後，よりWebページの評価文を抽出する手法を検討しなければならない．最後に，今回行わなかった関連リンク集の出力方法や掲載するページの選別方法についても検討する必要がある．

謝辞

本研究を進めるに当たり，熱心なご指導を賜りました白井清昭ならびに島津教授には，心から感謝を致します．多くのご助言，を頂きました山田寛康助手にに深く感謝いたします．さらに，自然言語処理学講座の皆様方には，貴重なご支援を頂きましたことを感謝いたします.

関連図書

[1] Satoshisato,Madoka Sato Automatic Generation of Directories for Specific Categories. In AAAI Workshop on Infomaation Systems,Orland,huly 18-19,1999.

[2] 板橋英夫，望月源，白井清昭，奥村学．参照ページからの情報を利用した Web 探索支援，第 8 回言語処理学会年次大会，pp.471-473,Mar.

[3] E.Amitay.IncoCommonSense-Rethinking Web Search Results. ICME 2000.

[4] 難波英嗣，奥村学．論文間の参照情報を考慮したサーベイ論文作成支援システムの開発，自然言語処理，Vol6,No5,pp43-pp62.

[5] 小林のぞみ，乾健太郎，松本祐治，立石健二，福島俊一．テキストマイニングによる評価表現の収集，情報処理学会研究報告，NL-154,pp.77-84.2003,Mar.

[6] 長江朋，望月源，白井清昭，島津明．製品コンセプトと製品評価文章の関係の分析，第 8 回言語処理学会年次大会，pp.583-586,2002,Mar.

[7] 村野誠治，佐藤理史．文型パターンを用いた主観的評価文の自動抽出，第 9 回言語処理学会年次大会，pp.67-70,2003,Mar.

[8] 小林のぞみ，乾孝司，乾健太郎，語釈文を利用した「p/n 辞書」の作成，人工知能学会言語・音声理解と対話処理研究会資料，SLUD,2001

[9] Ellen Riloff, Janyce Wiebe. Learning Extraction Pattern for Subjective Expression. EMNLP,2003