研究指導　中澤真　准教授

地域情報検索システムにおけるフィルタリングアルゴリズム

馬上　愛菜

1 はじめに

近年，インターネット上の情報量は膨大なものとなっており，webページの数も1兆ページにまで達している[1]．この膨大な情報の検索手段として，日本で最も利用されているサービスがYahoo!JAPAN[1]であり，検索エンジンとカテゴリ型検索の両方の機能を採用し，多様な検索方法を兼ね備えている[2]．

このようなサービスを利用して，ユーザは言葉の意味や定義，人物に関する情報，地域の施設や地名など様々な情報を検索している[3]．特に地域情報についてのニーズは高いため，検索エンジンでの地域に関する情報の検索や，地域に特化したカテゴリ型検索のサービスが増加している．しかし，これらのシステムにもまだ課題が多く残されている．まず検索エンジンの場合，網羅率[2]を活かして豊富な情報を提示できることが利点ではあるが，ランキング処理アルゴリズムの特性上，大型サイトがランキング上位を占める傾向にあり，ユーザの目的に合わせて地域固有のサイトを探すのが難しくなってしまう場合が多々ある．

そこで本研究では，これらの問題点を改善するために，既存の検索エンジンと同様にクローラを用いることで情報の網羅率･最新性[3]を維持しつつ，収集したサイトを特性別にフィルタリングすることで，ユーザの目的に応じた検索結果を提示できるシステムを提案する．今回はこのシステムの構成要素の1つである，サイトフィルタリングのアルゴリズムに重点を置き，サイトの特性をHTMLソース上の頻出または共起キーワードから判別する手法を提案する．そして，このフィルタリング手法の有効性を再現率･適合率によって示す．

2 情報検索システムの現状と問題点

現在，ユーザがインターネット上の情報を検索する際に利用するシステムは，検索エンジンとカテゴリ型検索の2つに分類される．ここでは，その2つの特性について論じる．

2.1 既存の情報検索システムの特性

まず検索エンジンとは，クローラと呼ばれるソフトウェアがネット上のサイトを自動収集･登録し，この情報に基づいてユーザからの検索要求に適したサイトを提示するシステムである．このシステムには2つの利点がある．1つ目は，クローラによる自動巡回により，常に最新の情報を収集することができること，2つ目は膨大な量のサイトを巡回することが可能であり，情報の網羅率が高いことである．例えば，代表的な検索エンジンであるGoogle[4]は高い網羅率を誇っており，現在登録されているwebページの数は80億を越えている［4]．しかし，これら膨大なサイトの中からユーザが求めるサイトを的確に出力するのは容易なことではないため，検索精度の問題が常につきまとう．実際に，得たいと思う情報が検索エンジンの結果として出力されないと考えているユーザも多い[5]．

一方カテゴリ型検索とは，インターネット上のサイトを手作業で収集し，カテゴリ別に分類した情報検索システムである．このシステムの利点は，手作業で情報を収集しているため，確実にカテゴリに適した情報を集約していることである．しかしこのシステムは，検索エンジンのようにクローラによる自動収集ではなく，手作業によるサイト収集を行っているため，情報の網羅率･最新性が低くなってしまうという問題がある．

2.2 地域情報検索における問題

次に前述した検索システムを利用して地域情報を検索した場合の問題点を明らかにする．

2.2.1 検索エンジンの問題

まず検索エンジンを用いて地域情報を取得した場合，どのような結果を表示するか検証した．実際にGoogleを用いて「福島県　会津若松市　ラーメン」というキーワードで検索し，上位100サイトに市外の情報が掲載されているか検証した．その結果，上位100サイト中2サイトのみが，市外に関する情報を掲載しており，他のサイトは市内の地域情報を掲載していた．このように，現在の検索エンジンは検索キーワードに地域名を指定することで，特定地域の情報を絞り込む能力を十分に有していることがわかる．

ただし，100サイト以上の検索結果が表示されても，

[1] http://www.yahoo.co.jp/
[2] インターネット上に存在する全webページからクローラがデータベースに登録したwebページの割合．
[3] 更新された情報が常に掲載されている状態．
[4] http://www.google.co.jp/

一般的なユーザは検索結果の1ページ目に表示される上位10サイト程度しか閲覧しないため[3]，ランキング処理の性能が非常に重要となる．そこで，この上位100サイトのランキングが的確か，また偏りがないかという点について調査した．その結果，施設情報を集約した地域または施設ポータルサイトが上位を占め，検索結果に偏りが生じていた．これらのポータルサイト内の情報は掲載店舗数が少数であったり，掲載店舗数は多いが各店舗に関する情報が住所･電話番号の基本情報のみであったりと，情報の網羅性や充実性について問題があった．一方，検索結果の下位にわずかに表示されていた施設が独自に運営するサイトや，個人のブログサイトであり，施設に関する詳細情報，あるいは施設を利用してのレビューなど情報が充実している．このように，サイトによって特性が異なるため利用目的に応じて使い分ける必要がある．そこで地域情報を検索するユーザの利用目的を2つに分けて考えることにする．1つ目は住所や電話番号，商品の種類や値段，アクセス方法など施設に関する詳細情報，2つ目は第三者からの施設に関する感想や評価の情報である．

この目的別に検索結果を表示するために，サイトを3つのタイプに分類することを考える．まず施設運営サイト[5]は施設側が情報を掲載するため情報の確実性が高く，施設に関する情報量が多いという特徴を持つため，ユーザが施設の詳細情報を得る目的に適している．次に，個人のブログサイトは施設の感想や商品の写真を掲載されているという特徴を持ち，ユーザが施設の評価情報を得る目的に適している．最後に地域またはポータルサイトは，多店舗の施設情報や口コミを掲載しているサイトであるが，前述した2つのサイトより詳細な情報を掲載していない場合があるため，施設運営サイトやブログが網羅していない情報を補う役目に適している．以上の3つのタイプに当てはまらないサイトをその他[6]に分類した．

このようなサイトの分類は，検索結果をユーザの利用目的に応じて出力するためには不可欠である．しかし，現在の検索エンジンはサイトをタイプ別に出力するようには設計されていないため，地域または施設ポータルサイトが上位100サイトの大半を占めてしまう(図1)．この原因はGoogleなどの検索エンジンがPageRankアルゴリズム[7]などを利用して，webページのリンク構造から各サイトの重要度を評価し，検索結果のランキング処理に反映させているためである[6]．ポータルサイトの多くは全国の施設や飲食店を網羅的に扱う大規模サイトであるため，このサイトを支持する被リンク数が多くPageRankが高くなり上位表示される傾向が強くなる．しかし個人事業主が運営しているような施設独自のサイトの場合，情報は充実しているが被リンク数が少なくPageRankが低いため，上位には表示されにくくなる．実際に，福島県会津若松市内のラーメン店が運営するサイトは12サイトあるにもかかわらず，検索結果として表示されたのは上位100サイト中3サイトのみであった．

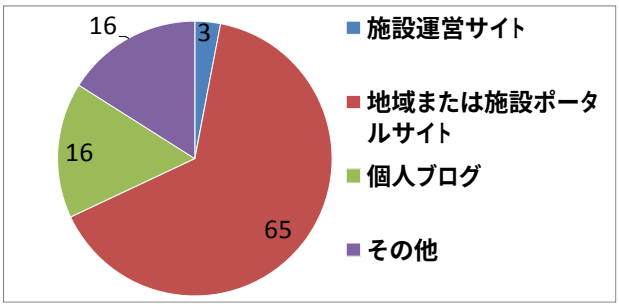

図 1 検索されたサイトの分類結果

2.2.2 地域に特化したカテゴリ型検索サイトの問題

一方検索エンジンを用いる以外に，カテゴリ型検索を用いて地域情報を検索する場合もある．例えば，全国の施設情報やイベント情報を掲載している「Yahoo!地域情報[8]」や対象範囲を会津若松に絞り，この地域の観光施設や飲食店情報を掲載している「会津若松観光物産協会公式サイト[9]」などが挙げられる．このようなサイトは，検索したい地域とカテゴリを指定することで，その地域の施設情報を的確に取得することができる．

しかし，これらのサイトにはカテゴリ型検索サイト特有の網羅性の問題がある．例えば，「iタウンページ[10]」上に掲載されている若松市内のラーメン店は74件あるが，「会津若松観光物産協会公式サイト」に掲載されているラーメン店の数は2件のみであり，網羅率が極めて低いことが分かる．一方，「Yahoo!地域情報」において同様の検索をした場合には，70件検索することができた．しかし，網羅率は高くとも，1店舗に関する情報は店名，住所，電話番号のみに限定されており，メニューや値段，アクセス方法などの詳細情報は掲載されていない．この

[5] 施設自身が店の概要や商品に関する情報を掲載し、独自に運営するサイト.
[6] ユーザにとってあまり必要としない情報が掲載されているサイト.
[7] コンピュータに仕事をさせるための手順.
[8] http://local.yahoo.co.jp/
[9] http://www.aizukanko.com/
[10] http://itp.ne.jp/

ように，掲載されている店舗数が少ないといった網羅率の低さ，店舗に関する情報の希薄さという点で検索エンジンの検索結果に劣ることが多い．

3　自動更新地域情報カテゴリ型検索の提案

3.1　システムの全体像

地域情報検索における既存システムの問題を解決するためには，カテゴリ型検索のようなカテゴリに分類され，かつ情報の網羅率･最新性を高めるために情報の自動収集機能を持った検索システムが必要である．そこで，このような要件を満たすシステムの全体像についてまず述べる．

このシステムは既存の検索エンジン同様，クローラによってサイトの収集を行う．次にフィルタリングによって，サイトをカテゴリ別に分類し，これをさらにサイトのタイプ別に分類する処理をする．これをデータベースに登録することで，ユーザの利用目的に応じた検索結果を出力する(図2)．本研究では，このシステムのサイトをタイプ別にフィルタリングする処理に焦点をあて，そのアルゴリズムについて提案する．

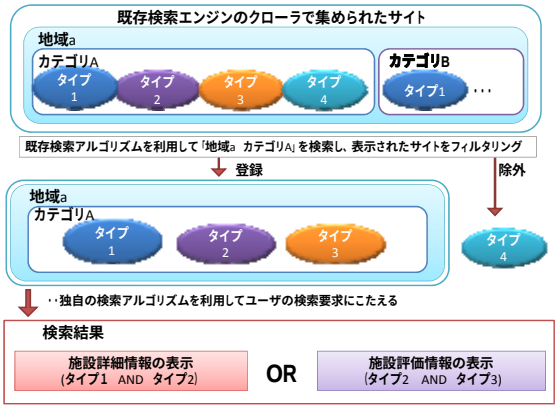

図 2　システムの全体図

3.2　サイトフィルタリング手法

クローラが巡回して取得したサイトの集合に対し，特定の地域や特定のカテゴリについて絞り込むためには，既存の検索エンジンのアルゴリズムで可能であることは既に述べた．ここでは，検索されたサイトをタイプ別に分類するために，サイトフィルタリングを多段階で実行するシステムについて説明する(図3)．これは，各タイプの判別をフィルタリングによって一つずつふるいにかける方式であり，第一段階でブログサイトであるかどうかの判別，第二段階で地域またはポータルサイトであるかどうかの判別，第三段階で施設運営サイトであるかどうかの判別を行い，最後まで分類されなかったサイトは除外する．

これらの各フィルタリングでは，各サイトに含まれるキーワードの有無で判別するキーワードフィルタリングを用いる．以下では，このキーワードの抽出方法と，そのキーワードを用いた判別方法について述べる．

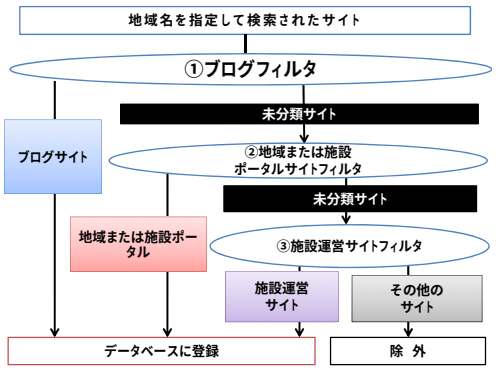

図 3　フィルタリングの手順

3.2.1　共起によるブログサイトのフィルタリング

第一段階のブログフィルタは，ブログタイプのサイトとそれ以外のサイトに判別するための役割を持っている．初めにブログサイトをフィルタリングする理由は，ブログにはブログ特有のキーワードが多く含まれているため，他のタイプとの識別がもっとも容易であるからである．

まずフィルタの判別ルールを構築するために，ブログ特有の語の抽出をする．今回は「福島県　会津若松市　ラーメン」のカテゴリにおいてサイトをタイプ別に分類することを想定し，このキーワードで検索した結果の上位100サイト中のブログタイプのサイトからキーワードの抽出をする．抽出の際に問題となるのが，コンテンツのどの部分からキーワードを抽出するかということであるが，ブログサイトの場合にはブログ特有のキーワードとなる機能名などがHTMLソース本文に含まれているため，ソース全文から抽出することにした．しかし，HTMLソース全文を対象とする場合，対象となるキーワード数が多く，頻度が高いキーワードを1語選んだだけでは他タイプにも頻出するものを抽出してしまう可能性があるため複数語の抽出が必要である．そこで，キーワードの選び方として共起関係に着目した．共起とは文書内にある単語が出現した際，別の単語が同時に出現することである．これらのキーワード間には，共起頻度が高いほど強い関連性があると判断することができる．つまり，ブログを表す強い関連性があるキーワードを複合することで，サイトがブログであるということをより正確に示すキーワードを抽出することができると考えた．今回の検証データでは，16のブログタイプサイトから抽出した共起頻度の高いキーワードは，「トラックバック　AND　ブログ」，「トラッ

クバック　AND　コメント」となり，ブログ特有の機能に関するキーワードが含まれることも確認できた．

最終的にブログフィルタは，前述した方法で抽出したキーワードを用いてブログの判別処理をする．具体的には，フィルタリング対象サイトのHTMLソース全文内に，抽出した共起キーワードのペアのいずれかが含まれていた場合，このサイトをブログタイプと判別しデータベースに登録する．

3.2.2 地域または施設ポータルサイトフィルタリング

第二段階の地域または施設ポータルサイトフィルタは，ポータルサイトに属するタイプとそれ以外に判別するための役割を持っている．このタイプのサイトはブログと違い特有のキーワードがないため，HTMLソース全文ではなく，トップページのHTMLソース内に含まれるmetaタグ[11]のひとつである，descriptionタグ内に記述されている文章から抽出する．descriptionタグとは，サイトの説明文を記載する部分であり，検索エンジンにサイトの内容を理解させるとともに，検索結果のスニペット[12]として利用されることが多い．それゆえ，このタグ内には地域または施設ポータルサイト特有のキーワードが使用されている可能性が高いと考え，キーワードの抽出対象とした．提案するフィルタでは，このタイプのサイトの集合に対して，多くのサイトで生起する頻出キーワードをdescriptionタグ内に限定して抽出する．

ただし，他タイプに属するサイトにも共通して出現率の高いキーワードは除外する必要がある．なぜなら，このようなキーワードは，他のタイプに属すサイトを誤判別する原因になりやすいからである．例えば今回の検証データの場合，「ラーメン」というキーワードは他タイプのサイトにおいても出現頻度が高いため，このキーワードは抽出対象から外す(表1)．

さらに，フィルタリング対象のHTMLソース内にdescriptionタグが存在しないことによるフィルタリング漏れを防ぐために，ブログフィルタと同様のソース全体に対する共起キーワードも予備的に抽出しておく．これにより，descriptionタグ内のキーワードフィルタリングよりも精度は落ちるが，フィルタリング漏れの確率をある程度抑えることが可能になる．

このフィルタでは，ブログフィルタによってブログタイプと判断されなかったサイトを対象としてフィルタリング処理をする．具体的にはフィルタリング対象サイトのdescriptionタグ内に，先ほど抽出した頻出キーワードのいずれかが含まれている場合には，地域または施設ポータルサイトタイプとして判別し，データベースに登録する．また，descriptionタグを持たないサイトの場合は，HTMLソース全体に対して，抽出済みの共起頻度の高いキーワードの含有の有無によって判別する．

表 1　地域・施設ポータルの頻出キーワード[13]

順位	頻出キーワード	地域または施設ポータルサイト	施設運営サイト	その他
1	情報	37/65	0/12	1/12
2	検索	24/65	0/12	0/12
3	口コミ OR クチコミ	25/65	0/12	1/12
4	ラーメン	10/65	8/12	1/12
5	ポータル	9/65	0/12	1/12

3.3.3 title をもとに施設運営サイトのフィルタリング

第三段階の施設運営サイトフィルタは，施設運営サイトタイプと，その他のタイプに判別するための役割を持っている．このタイプの場合のHTMLソースに含まれるキーワードは，他のタイプにも共通するキーワードが多く，タイプ分けすることが難しいと考えられる．ただし，施設運営サイトにはそのサイト内のいずれかのHTMLソースに施設の住所が掲載されているという特性がある．そこで，この性質を利用することにより，サイト内の住所表記とiタウンページの該当施設の住所情報を照合して施設名を特定し，この名称がHTMLのソース内で用いられているかどうかでフィルタリング処理をする．

ただし，施設名称の照合は，HTMLソース全体に対して行うのではなく，トップページのtitleタグ内の文章に対して行うべきである．なぜなら，titleタグはサイト名が記載される場所であるため，施設運営サイトの場合には店舗名が含まれている確率が非常に高いからである．

以上のように段階的にフィルタリングを施し，最終的に「その他のサイト」として分類されたサイトは，データベースへの登録をせずに除外することになる．

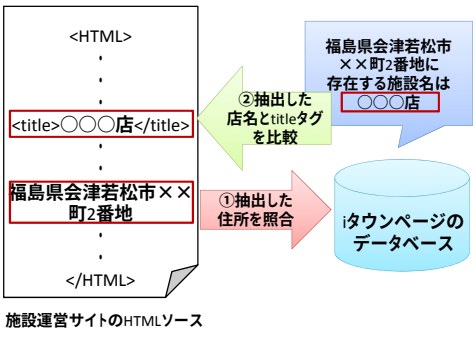

図 4　title タグを用いたフィルタリング手法

[11] ソースに記載される文書の情報をブラウザやクローラに知らせるためのタグ．

[12] 検索結果のページタイトルの下に表示される説明文．

[13] 「口コミ　OR　クチコミ」は言葉の揺らぎを考慮している．

3.4 サイトフィルタリングの検証結果

最後に,「福島県　会津若松市　ラーメン」というキーワードをGoogleで検索した上位100サイトを対象として, 提案したフィルタリングの有効性を検証する. なお, この検索結果の上位100サイトには施設運営サイトが3サイトしかふくまれないため, ランキング外の9サイトも追加して検証することにした. ここではフィルタリングの性能の評価基準として, 再現率[14]・適合率[15][7]を用いる.

まず再現率とは, 抽出すべきサイトに対し, フィルタリングによって正確に抽出されたサイトがどのくらい占めているかを割合で表したものである. 結果として, すべてのフィルタリング手法の再現率が90%以上の成果を上げることができた(表2).

次に適合率とは, フィルタリングによって抽出されたサイト内にどのくらいの抽出すべきサイトが占めているかを割合で表したものである. これにより, 誤って他のタイプに分類されるべきサイトを抽出していないかを検証することができる. 結果として, すべての手法の適合率は約89%以上の成果を上げることができた(表3).

表 2　フィルタリング効果測定（再現率）[16]

フィルタの種類	再現率	再現率計算式
ブログフィルタ	100.0%	=16/16
地域または施設ポータルサイトフィルタ	90.5%	=(52+5)/(57+6)
施設運営サイトフィルタ	91.7%	=11/12

表 3　フィルタリング効果測定(適合率)

フィルタの種類	適合率	適合率計算式
ブログフィルタ	88.9%	=16/18
地域または施設ポータルサイトフィルタ	96.6%	=(52+5)/(54+5)
施設運営サイトフィルタ	100.0%	=11/11

4 むすび

本研究では, 地域情報を既存検索システムで検索した際, ユーザの目的に応じたサイトが検索されないことを問題として取り上げ, 改善策として検索エンジンとカテゴリ型検索の利点を活かし, ユーザの目的別にサイトを表示する検索システムの提案をした. 今回はこのシステムの構成要素の1つであるサイトフィルタリング手法を提案し, その有効性を検証した.

その結果, 提案したすべてのフィルタリング手法が再現率・適合率ともに約89%以上となり, 高い精度でサイトフィルタリングを実現することができた. しかし今回フィルタリング対象とした地域情報に偏りがあるため, 他の地域情報の有効性についても確かめる必要がある. また, 今回はシステムの構成要素の1つである, サイトのフィルタリング部分のみの提案となってしまった. そのため, システムを構築するためには, クローラによるサイトの巡回アルゴリズムやデータベースに登録されたサイトを, どのような手法でランキングするべきか検討を続けていく必要がある.

[14] 再現率=正解である検索結果の数/全正解数.
[15] 適合率=正解である検索結果の数/検索結果の数.
[16] 計算式の括弧内は, description タグを使用したサイトと HTML ソース全文を使用したサイトの数を足し合わせている.

参考文献

[1] OfficialGoogleBlog, We Knew The web was big, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

[2] comSCORE, Yahoo! Attracts More than Half of All Searches Conducted in Japan in January 2009, http://www.comscore.com/jpn/Press_Events/Press_Releases/2009/3/Japan_Search_Engine_Rankings

[3] 中村聡史, “情報検索に対する信頼性調査および結果, ”人口知能学誌23巻6号, p.769, 2008.

[4] 西田圭介, Googleを支える技術　巨大システムの内側の世界, 技術評論社, 2008.

[5] Japan.internet.com, 検索エンジン、約2割が「キーワードと検索結果が一致しない」ことに不満, http://japan.internet.com/wmnews/20080416/2.html

[6] AmyN・Langville, CarlD.Meyer, Google　PageRankの数理　最強エンジンのランキング手法を求めて, 共立出版, 2009.

[7] 河野浩之, 山田誠二, 北村康彦, 高橋克己, 情報検索とエージェント, 学校法人　東京電機大学, 2002.