語義の曖昧性を考慮したWWW関連リンク集の自動生成

田村　雅樹　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{m-tamura, kshirai}@jaist.ac.jp

1 はじめに

我々は，ウェブでの情報探索を支援することを目的に，あるテーマが与えられたとき，そのテーマに関する情報をウェブから収集し，ポータルサイトを自動的に構築する研究に取り組んでいる [2]．ポータルサイトとは，一般には Yahoo のような総合ディレクトリサービスを指すが，ここではユーザがあるテーマに関する情報を調べるときに最初に訪れる入口となるべきサイトを指す．本研究は，ポータルサイトのコンテンツのひとつとして，あるテーマがいくつかのキーワードとして与えられたとき，そのテーマに関するページを自動的に収集し，関連リンク集を自動的に生成する手法を提案する [3]．その際，以下の 2 点に特に留意する．

- キーワードの意味の曖昧性
  テーマとして与えられたキーワードがいくつかの意味を持つとき，その意味毎に関連するウェブページを集めてリンク集を作成する．例えば,「松井」に関するリンク集を自動作成する場合を考える.「松井」には「松井秀喜」や「松井稼頭央」といった何人かの人物がいるが，1 つのリンク集に「松井秀喜」や「松井稼頭央」に関するページが混在するのは望ましくない．そこで，キーワードの意味の曖昧性を自動的に判別し，その意味毎に関連するページを集めてリンク集を作成する．
- 既存のリンク集に対する処理
  本研究では，自動構築するリンク集の中には既存のリンク集を含めない．メタリンク集も情報探索にとって有効であるが，目的のページに到達するまでに 2 回以上リンクを辿ることになり，好ましくない．テーマに関連する既存のリンク集を見つけたら，そのリンク集に掲載されているページを候補ページに加え，新たにリンク集を自動構築する．すなわち，リンク集自動構築の処理には既存のリンク集の再編も含まれる．

2 提案手法

本節では，あるテーマが与えられたとき，そのテーマに関連するウェブページを収集し，関連リンク集を生成する手法について述べる．ここで，テーマはいくつかのキーワード (名詞) で与えられるものとする．キーワードの集合を K，個々のキーワードを k_i とする．

2.1 候補ページの取得

キーワード k_i をクエリとしてウェブ検索を行う．検索エンジンとして goo [1]を用いた．検索結果として得られた上位のページを 500 件を上限として取得し，リンク集に掲載すべきページの候補とする．以下，候補ページの集合を P とする．

2.2 既存リンク集の処理

候補ページの集合 P に既存のリンク集が含まれている場合を考える．もしこのリンク集がテーマに関連しているとすれば，そのリンク集に掲載されているページはシステムが自動構築するリンク集にも加えるべきページである可能性が高い．そこで，以下の手続きにしたがって候補ページの追加を行う．

1. P の中から既存のリンク集を検出する．リンク集を検出する方法については後述する．
2. 検出したリンク集の領域の中にキーワード k_i が 1 個も含まれていなければ，そのリンク集はテーマとの関連性がないとみなし，3. の処理は行わない．
3. それ以外のリンク集については，リンク集に掲載されているリンクを辿り，ページを取得する．そのページが全てのキーワードを含んでいるのなら，テーマへの関連があるとみなし，P に追加する．ただし，既に P に含まれているページは除く．

また，P の中にリンク集だけからなるページがあれば，それを候補ページから除去する．なぜなら，リンク集に別のリンク集を入れると，ユーザは目的のページに到達するまでに 2 回リンクをクリックすることになるので好ましくないからである．また，既に上記の操作で既存のリンク集に掲載されているページは P に含まれているはずなので，リンク集自体をあえて P に残す必要はない．ここでは，リンク集として検出した領域がページ全体の 80%以上のときに，そのページの主なコンテンツ

[1] http://www.goo.ne.jp/

はリンク集のみであると判断して P から除去する．ただし，この判定はタグを除いたテキスト部分のみで行う．

既存のリンク集の検出

リンク集の検出は HTML タグをもとにしたパタンマッチにより行う．以下のいずれかのパタンにマッチする領域をリンク集と判定する．

- ul タグが囲む範囲の中で，li タグの直後に外部リンクが 3 個以上あるとき，ul タグが囲む範囲をリンク集とする
- 外部リンクの後に br タグが続くパタンが 3 回以上出現するとき，その外部リンクの集合をリンク集とする．ただし，外部リンク (a タグ) と br タグの間にはインライン要素のタグ (b,font など) 以外のタグが現われてはならない．
- table タグが囲む範囲で，外部リンクが存在する行が 3 個以上あるとき，table タグが囲む範囲をリンク集とする．

ここで外部リンクとは，別の作者によるページへのリンクであり，同一作者のページへ張るサイト内リンクと区別している．あるリンクが外部リンクであるかどうかの判定は以下のように行う．

- href 属性の値が相対パスなら外部リンクではない．
- ページ自身の URL とリンク先ページの URL を比較し，ドメインとその直下のディレクトリ (例えば以下の下線部) が同じなら外部リンクではない．

  http://www.xxx.com/aaa/bbb/index.html

- 上記以外は外部リンクであるとみなす．

2.3 クラスタリング

収集した候補ページの集合 P に対してクラスタリングを行い，いくつかのクラスタを作成する．通常の文書クラスタリングでは，同じトピックに関する文書をひとつのクラスタにまとめる場合が多い．これに対し，ここでのクラスタリングはキーワードの意味の曖昧性に基づくものであり，同じ意味で使われるキーワードを含むページをまとめて 1 つのクラスタを作る．

2.3.1 キーワードの意味の決定

P 中のページに出現するキーワード k_i について，その前後に出現する名詞に着目し，(p, k_i, s) という形式で抽出する．ここで，p は k_i の直前に現われて k_i とともに複合名詞を構成する語，s は k_i の直後に現われて複合名詞を構成する語である．前後に何もない場合は，p または s を空列 ε とする．

キーワードの前後に現われる名詞はその名詞の意味の曖昧性を表わすとみなせる．例えば，キーワードが「松井」のとき，$(\varepsilon, \text{松井}, \text{秀喜})$ や $(\varepsilon, \text{松井}, \text{稼頭央})$ はそれぞれ人物の違いを，キーワードが「野球」のとき，$(\text{高校}, \text{野球}, \varepsilon)$ や $(\text{プロ}, \text{野球}, \varepsilon)$ は野球の種類の違いを表わす．

次に，各候補ページに対して，そのキーワードの代表的な意味を決定する．そのページ内における (p, k_i, s) の出現頻度を調べ，出現頻度の最も大きい (p, k_i, s) をひとつ選択する．すなわち，k_i はそのページでは (p, k_i, s) の意味を持つとみなす．また，出現頻度が 2 位以降の (p, k_i, s) のうち，一位の出現頻度と比べて 75%以上の頻度を持つものは，その (p, k_i, s) もやはりそのページにおける k_i の意味とみなす．これは，1 つのページに複数のトピックが記述され，k_i が複数の意味として使われている可能性を考慮したためである．

2.3.2 基本クラスタの作成

上記の操作で，P 中の各ページについて，キーワード k_i の代表的な意味 (p, k_i, s) がいくつか決まる．次に，個々の (p, k_i, s) に着目し，(p, k_i, s) をキーワードの代表的な意味として含むページの集合を求め，基本クラスタ $bc(p, k_i, s)$ とする．例えば，$bc(\varepsilon, \text{松井}, \text{秀喜})$ は $(\varepsilon, \text{松井}, \text{秀喜})$ を「松井」の代表的な意味として持つページの集合である．ただし，$bc(\varepsilon, k_i, \varepsilon)$ という基本クラスタは作成しない．キーワードの前後に何もない場合，そのキーワードがどのような意味で使われているか不明であるからである．

最終的には個々の基本クラスタが 1 つのリンク集になるが，あまりにページ数が少ない基本クラスタはリンク集としてふさわしくない．そこで，得られた基本クラスタをクラスタに含まれるページ数の降順にソートする．そして，上位 5 位以上であり，かつページ数が一位の基本クラスタのページ数の 10%以上である基本クラスタのみを残し，これらを生成するリンク集の候補とする．

また，基本クラスタに属さないページをまとめて「その他」クラスタ oc を作る．

2.3.3 基本クラスタへのページの追加

たとえ前後に現われる名詞が異なる場合でも，キーワードの意味としては同じになる場合がある．例えば，「松井」というキーワードが $(\varepsilon, \text{松井}, \varepsilon)$ というように単独で出現しても，前後の文脈から「松井秀喜」を指すとわかることもある．また，$(\text{ヤンキース}, \text{松井}, \text{秀喜})$ という形式で出現する「松井」は「松井秀喜」を指すとわかるが，2.3.2 で述べた操作ではこのページは $bc(\varepsilon, \text{松井}, \text{秀喜})$ には属さない．ここでは，特に $(\varepsilon, k_i, \varepsilon)$ という

ように単独で現われるキーワードの意味を特定するために，基本クラスタへは未分類のままである oc クラスタ内の個々のページと，基本クラスタとの類似度を計算し，類似度が高い場合にはそのページを基本クラスタに追加する．

クラスタとページの類似度計算は単語ベクトルを基準に行う．まず，ページ p の単語ベクトル $\vec{p}$ を式 (1) のように定義する．

$$\vec{p} = (w_1, \cdots, w_n) \tag{1}$$

w_i は p に含まれる単語 t_i に対する重みである．ここではページ中の全ての単語ではなく，キーワード k_i の前後 50 語以内に現われる自立語のみを単語ベクトルの要素とする．w_i は式 (2) のように定義する．

$$w_i = tf(p, t_i) \cdot icf(t_i) \tag{2}$$

$$icf(t_i) = \log\left(\frac{N_c}{cf(t_i)} + 1\right) \tag{3}$$

$tf(p, t_i)$ は単語 t_i のページ p における出現頻度である．ここでは頻度そのものではなく相対出現頻度とする．すなわち，$\sum_i tf(p, t_i) = 1$ とする．一方，$icf(t_i)$ は Inverted Cluster Frequency であり，情報検索でよく採用される IDF と同様に，単語 t_i がある特定のクラスタのみに現われるときに高い重みを与える働きをする．$icf(t_i)$ は式 (3) で定義される．$cf(t_i)$ は t_i が出現する基本クラスタまたは「その他」クラスタ oc の数であり，N_c は oc を含む基本クラスタの総数である．

一方，基本クラスタベクトル $\vec{bc}$ は式 (4) のように基本クラスタに属するページの単語ベクトルの和と定義する．

$$\vec{bc} = \sum_{p \in bc} \vec{p} \tag{4}$$

基本クラスタ bc と oc に属するページ p の全ての組み合わせについて，両者の単語ベクトル $\vec{bc}$ と $\vec{p}$ の類似度を計算し，ある一定の閾値を越えたとき，そのページを基本クラスタに追加する．ベクトル間の類似度は cosine 類似度を用い，閾値は 0.4 に設定した．また，計算の効率化のため，あるページを基本クラスタに追加後，別のページとの類似度を計算する前には基本クラスタの単語ベクトルは更新しない．基本クラスタ $bc(s, k_i, p)$ にページ追加した後のクラスタを $c(s, k_i, p)$ とする．

2.4 リンク集の選別

ユーザがテーマとしてキーワードを指定した際，どのような意味を想定してそのキーワードを指定したのかを推測することは難しい．そこで，前項までで得られた個々のクラスタを 1 つのリンク集の候補とし，それらをユーザに提示する．その際，クラスタが $c(s, k_i, p)$ なら，s,k_i,p を連結した文字列 (「松井秀喜」「松井稼頭央」など) を提示し，提示された複数のリンク集の違いが容易にわかるようにする．ユーザに適切なリンク集を選んでもらい，これを最終的に出力するリンク集とする．

3 予備実験

提案手法の評価を行った．表 1 に挙げた 5 つのテーマに対し，2.1〜2.3 項で述べた手法にしたがって複数のクラスタ (リンク集) を作成した．

表 1: テーマ一覧

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 }

まず，2.2 項で述べたリンク集検出アルゴリズムを評価した．候補ページの集合 P に対して既存のリンク集の検出を試み，1 つのテーマにつき，既存リンク集を検出したページを 15，検出しなかったページを 15，それぞれランダムに取得した．それらのページに対する既存リンク集検出の正解率を調べたところ，精度 (適合率) が 67.7%，再現率が 55.1%となった．特に再現率が低いのは，リンク集検出のためのパタンの不足が原因であった．本研究ではリンク集検出については精度よりも再現率を重視しているので，パタンの追加による再現率の向上は急務である．

次に，作成された基本クラスタを表 2 に示す．(ε, 松井, 秀喜), (プロ, 野球, ε) のようにキーワードの意味をうまく反映したクラスタもあれば，そうでないものもある．

また，個々のクラスタの中から初期の検索で得られたページを 15，既存のリンク集を辿ることによって候補に追加したページ (2.2 項参照) を 15 個ランダムに選び，それぞれテーマに沿ったページであるかを人手で判定した．結果を表 3 に示す．表 3 の「初期」は最初の検索エンジンによる検索で得られたページを,「追加」は既存リンク集を辿ることによって追加されたページを表わす．表中の数値はリンク集の掲載ページとしてふさわしいと判断されたページの割合であり，また括弧内の数値はページ数の合計である．全体的に見れば，リンク集に掲載するのに適したページの割合は 43%程度である．しかし，本研究はキーワードの意味の曖昧性に着目し，意味毎に別々のリンク集を作成することに焦点を当てており，現時点では個々のページがリンク集に掲載すべきかどうかの判定をほとんど行っていない．これは今後の重

表 2: 作成された基本クラスタ

1: (ε, 松井, 秀喜), (ε, 松井, 稼), (ε, 松井, 雄飛)
2: (ε, 石川, 県), (ε, 石川, テレビ), (ケーブル, テレビ, ε), (ε, テレビ, 番組表), (ε, テレビ, 番組), (テレビ, 番組表, ε), (週間, 番組表, ε)
3: (*perl*5, リファレンス, ε), (ポケット, リファレンス, ε), (*perl*, リファレンス, ε)
4: (日本, 地図, ε), (ε, 日本, 地図), (ε, 日本, 全国)
5: (高校, 野球, ε), (プロ, 野球, ニュース), (プロ, 野球, ε), (高校, 野球, 部)

表 3: リンク集の評価

初期	追加	合計
0.494 (657)	0.345 (531)	0.426(1,118)

要な課題と考えている．一方，リンク集を辿ることによってページを追加したことの効果については，追加されたページ数は多いものの，初期の検索で得られたページと比べて不適切なページが多く含まれることがわかる．とはいえ，テーマによっては適切なページを数多く追加したケースもあり，リンク集を辿ることの効果はある程度認められる．

最後に，本研究で提案するクラスタリング手法の評価を行った．実験の結果，それぞれの基本クラスタに対し，類似度計算によって基本クラスタに新たに追加されたページ数の合計は 925 であった．それらのページについて，そのページ中のキーワードが基本クラスタが定義するキーワードと同じ意味を持つかどうかを人手で判定した．その結果，約 50%のページが基本クラスタを定義するキーワードと同じ意味を持っていた．ただし，クラスタに対する正解率の変動が大きく，正解率が 90%近いクラスタもあった．今後，単語ベクトルの重み付けや単語ベクトルに登録するべき単語の選択方法などを十分検討する必要がある．

4 関連研究

ウェブのリンク集を自動的に構築する試みとしては Sato らによるウェブディレクトリ自動構築が挙げられる [1]．Sato らは，水族館，動物園などに対して，住所・電話番号などの情報をウェブから収集し，統合・再編した上でひとつの住所録を作成している．しかし，本研究のようにキーワードの意味の曖昧性は特に考慮されていない．一方，白井ら [2] は，関連リンク集を作成することを目的に，リンク集に掲載するべきページの説明や第三者による評価などをウェブから獲得することを試みている．しかし，本研究のようにリンク集に掲載すべきページの選別は行っていない．

ウェブページのクラスタリングに関する研究もいくつか行われている [5] が，その多くは文書のトピックに着目してクラスタリングを行っている．これに対し本研究では，同じ意味を表わすキーワードを含むページを 1 つにまとめることを目的にクラスタリングを行っている．本研究のクラスタリングと近い立場にあるのは検索エンジン Vivisimo[4] である．Vivisimo は，クエリを与えると，検索結果を自動的にクラスタリングしてユーザに提示する．クラスタリングの基準は，本研究と同様に，キーワードの前後の出現する名詞に着目し，その出現する名詞毎にウェブページを選別しているようである．このシステムの詳細は不明だが，本研究のクラスタリングのように，キーワードの前後の文脈が違う場合でも同じ意味を持つキーワードを含むページをひと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．

5 おわりに

本研究では，ポータルサイトのコンテンツとして関連リンク集を自動構築する手法を提案した．特に，キーワードの意味の曖昧性を考慮し，意味別に関連リンク集を作成すること，既存のリンク集を検出したとき，それに掲載されているページを候補ページに追加し，リンク集の再編を行う点に特徴がある．現在のところ実験結果が総じて悪いので，手法をさらに洗練し，実用に耐えうるシステムを構築したいと考えている．特に，現状のシステムは個々のページの内容をそれほど厳密にチェックしてリンク集を構築しているわけではない．リンク集に掲載すべきページとは何か，どのような特徴を持っているのかをよく検討し，適切なリンク集を作成できるようにしたい．

参考文献

[1] Satoshi Sato and Madoka Sato. Automatic generation of web directories for specific categories. In *AAAI Workshop on Intelligent Information Systems*, 1999.

[2] 白井清昭, 菅井俊介, 平野健児, 星正人. ポータルサイト自動作成の試み. 第 10 回言語処理学会年次大会, pp. 624–627, 2004.

[3] 田村雅樹. WWW における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, 3 2006.

[4] Vivisimo. `http://vivisimo.com/`.

[5] Oren Zamir and Oren Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 46–54, 1998.