L-018

検索エンジンからのメタ情報創出に関する研究

Research about Creating Metadata from Search Engine

池辺　正典　†　田中　成典　‡　中村　健二　†　古田　均　‡
Masanori Ikebe　Shigenori Tanaka　Kenji Nakamura　Futoshi Furuta

1.　はじめに

近年，インターネットの普及に伴い，日本でも ADSL や光ファイバーなどブロードバンドの高速通信網のインフラ（情報基盤）が本格的に整備され，インターネット利用者が爆発的に増加した[1]．また，インターネットの情報量は，年々増加の一途をたどり，目的のページを探し出すために多大な労力が必要となってきた．そのため，インターネット利用者の内，約 8 割以上の人が検索エンジンを利用して，目的の Web ページを探している．

Web ページの作成者は，アクセス数を増加させるための方法として検索エンジン最適化（SEO）対策[2]を日々行っているのが現状である．具体的には，任意の Web ページの構成を変更し，検索エンジン（検索用ロボット）が解析しやすいように変更し，Web ページを検索エンジンに最適化することを目的とするものである．

検索エンジンの検索結果は，利用者が入力したキーワードに関係が深い情報を表示する仕組みになっている．そのため，利用者が得ることのできる情報は，特定のキーワードに関して検索エンジンが自動的に割り出した順位情報[3]のみである．しかし，検索エンジンの検索結果に表示される情報は，検索エンジンが保持する情報の一部にしか過ぎない．そこで，検索結果を算出する過程で発生する情報を有効活用することにより，新たな知識を生み出すことができる．具体的には，SEO 対策や自サイトの意味付けなど，様々な用途に応用することができる．そこで，本研究では，検索エンジンの検索結果を算出する過程の情報を推定し，メタ情報を創出することを目的とする．

2．　要素技術の概要

2.1　ロボット型検索エンジンのアルゴリズム

ロボット型検索エンジンは，複数のアルゴリズムの算出結果を基に検索結果順位を決定する．現在公開されている検索エンジンでは，検索エンジンスパムにより検索結果の精度が低下している．代表的な検索エンジンスパムとしては，Web ページ内に見えない文字で検索キーワードを記述して，検索結果の順位を向上する方法などがある．そのため，検索結果の精度を向上するために，対象の Web ページ内の情報のみではなく，ページ外要因も評価基準として利用している．ページ外要因の評価を行う，ロボット型検索エンジンのアルゴリズムの中で代表的なページ外要因の評価法を次に示す．

（1）リンクポピュラリティ

リンクポピュラリティとは，リンクが多いページは，有益なページであるという前提の基にページのランク付けを行うものである．Google の PageRank などは，リンクポピュラリティを基に算出されている．

（2）リンクレピュテーションテーション

リンクレピュテーションとは，特定のページからのリンクされている場合に，どういった評価をされているかを算定する手法である．リンクレピュテーションの概要を図１に示す．

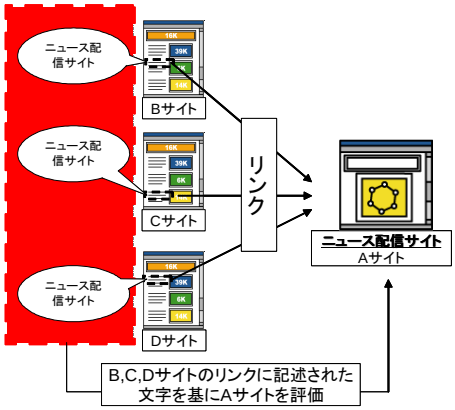

図１　リンクレピュテーションの概要

リンクレピュテーションでは，他サイトのリンクに表示されるアンカーテキストを基にサイトの評価を行う．図１では，B サイト，C サイトと D サイトは，A サイトにリンクしており，A サイトへのリンクのアンカーには，全て「ニュース配信サイト」と記述されているため，A サイトは，ニュース配信サイトとして登録される．

（3）クリックポピュラリティ

クリックポピュラリティとは，検索結果のページにおいて，「検索結果のリンクがクリックされた回数の多いページが優良なページである」という前提の基にページの表示順位を設定しているものである．

2.2　HTMLDOM

通常，HTML や XML などのタグ文書を扱うアプリケーションを開発する場合には，DOM（Document Object Model）と呼ばれる構造化文書解析用の API を利用するのが一般的である．特に，最近では XML 文書を構文解析・操作するための「XML DOM」が，オープンソースベースから商用ベースまで数多く開発されている．そして，HTML にも，構文解析・操作するための API として HTML DOM が存在する．これらの DOM は，W3C（World Wide Web Consortium）が提唱するインターフェイスを持つ．HTML DOM を利用することにより，プログラムから HTML の構造や内容を抽出し，その内容を編集することが

†　関西大学大学院 総合情報研究科
‡　関西大学総合情報学部

容易になる． DOM は，HTML 文書を読み込むと，メモリ空間上に，DOM tree と呼ばれる木構造のオブジェクト群を生成する．これらのオブジェクト群は，HTML 文書の各要素に対応しており，アプリケーションは，DOM tree のオブジェクトに用意されたメソッドを呼び出すことにより，各要素の参照，編集，削除や追加を行うことができる．HTML DOM の概念を図２に示す．

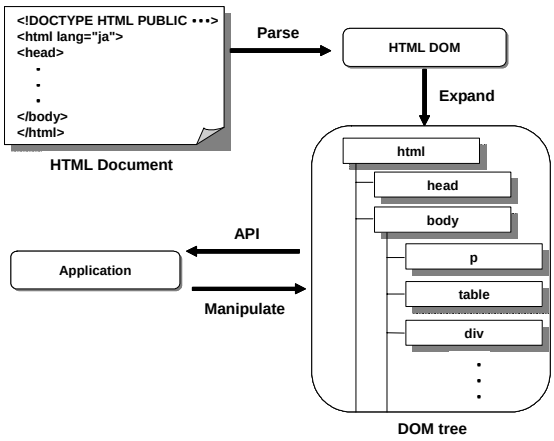

図２　HTML DOM の概念

HTML DOM を利用することにより，HTML の構文解析に必要な処理の多くを自動化できる．例えば，HTML DOM では，HTML 内の<A> 要素の href 属性だけを抽出するなどの処理が数ステップ程度の短いコードで実現できる．そのため，ページ内のデッドリンクを調査する「デッドリンクチェッカー」や Web サイトの「自動巡回ロボット」，「アクセシビリティチェッカー」など，HTML を解析するプログラムに幅広く HTML DOM の技術は利用されている．

2.3　FOAF

FOAF(Friend of a Friend)[4]とは，RDF/XML[5]によって記述された人と人との関係を表現するモデルである．FOAF は，RDF の今後の発展性を追及する RDFWeb の研究の一貫として発明された．また，人と人とのつながりを表現することにより，Web 上の属性や関連性をエージェントが解析できるようになり，機械が自動的に人と人とのつながりを理解できる．

FOAF は，情報の信頼性の向上や電子署名などセキュリティ面での活用方法も視野に入れており，セマンティック Web[6]を実現する上で必要不可欠な要素である「信頼性」を高める上で大きな役割を担う．

3.　研究の概要

本研究では，検索エンジンが表示順位を決定する過程でそぎ落とされる情報から，有用なメタデータを取得することを目的とする．ロボット型検索エンジンの中でも一般的な Google の順位算出[4]の流れを図３に示す．

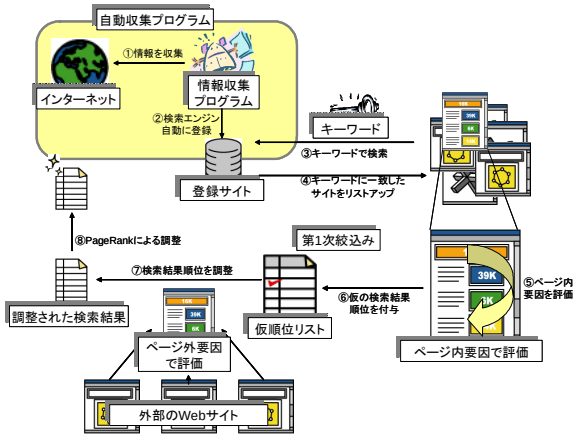

図３　Google の順位算出の流れ

Google の検索エンジン[7]の順位は，主に図３に示す過程に基づいて算出[8]される．まず，一定期間ごとに一回インターネットを自動巡回し，Web サイトの情報をデータベースに蓄積する．次に，利用者が検索した場合に，入力された検索キーワードに一致した Web サイトをリストアップする．そして，リストアップした Web サイトについて，ページ内要因を判別し，仮順位を設定する．加えて，他サイトの情報を基に仮順位を調節する．最後に，ページランクを考慮して，最終的な順位を決定する．

本研究では，検索エンジンの検索結果を算出する上で，ブラックボックスとなっているページ外要因の評価形式について調査し，検索結果を算出する過程で利用される情報を取得する．また，取得した情報が実際に優位な情報であるかを確認するために，SEO 対策をキーワードとして Web サイトの管理および更新を行う．

4.　システム概要

本研究では，検索エンジンの検索結果の有効活用性を確認するための方法論を提案する．本提案では，検索結果からメタ情報の創出するための機能として，検索エンジンの検索結果を解析し，Web リンクマップ，キーワードの有効性の算出と検索結果順位の取得機能を実現する．

4.1　Web リンクマップ

Web リンクマップでは，特定の Web サイトにリンクしているサイトを抽出し，各サイトの有効性の高いリンクを強調表示する．Web リンクマップの概念を図４に示す．

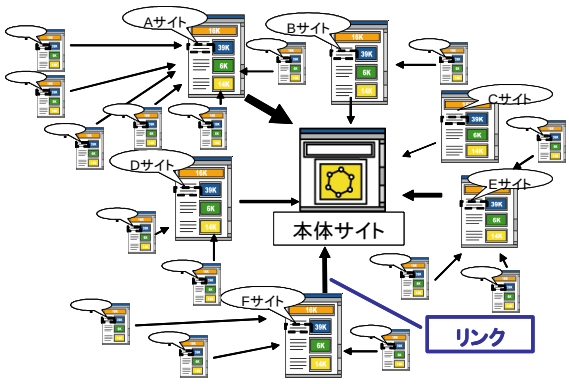

図４　Web リンクマップの概念

Web リンクマップは，次の手順で作成する．

- 本サイトにリンクしている Web サイトの一覧を取得（A サイト群）
- A サイト群の中の各 Web サイトにリンクして Web サイトの一覧を取得（B サイト群）
- A サイト群の中の各 Web サイトにリンクしている B サイト群の数を基に A サイト群の評価を判定
- A サイト群の評価を基に，自サイトへの貢献度を判定し，貢献度の比率により矢印の太さを調節

Web リンクマップでは，本体サイトへの貢献度が高いリンクを太い矢印で表示する．Web リンクマップの考え方は，Google のページランクを算出する過程で用いられるアルゴリズムである．Web リンクマップを形成することにより，本サイトへの貢献度のみではなく，対象サイトの信頼度を測定することができる．

今後は，本機能で生成された Web リンクマップをセマンティック Web へ応用することにより，自動的に FOAF のフォーマットに対応した RDF/XML で記述されたファイルを出力する．

4.2 キーワードの有効性の算出

キーワードの有効性の算出機能では，検索キーワードとしての有効性を算出する．キーワードの有効性は，Web サイトを作成する場合に一般的に利用され，検索キーワードとしても利用頻度が高いものである．各キーワードの有効性は，以下の方法で算出する．

- 有効性を算出するキーワードのリストを作成する．
- キーワードリストのキーワード毎に各検索エンジンにおける仮の有効度を算出する．有効度は，「検索エンジンのインデックスサイズ数／検索結果総数」とする．
- 検索エンジンから自サイトへ訪れる利用者が利用するキーワードと比較し有効度を調節する．

上記の手順から得られた有効度を基に，検索エンジン毎のキーワードの有効性を算出する．キーワードの有効性を算出することにより，本サイトに対して，利用者がどのような情報を期待しているかを予測することができる．そのため，SEO 対策の方向性の決定や事業化計画の立案に役立てることができる．

4.3 検索結果順位分析

検索結果順位分析機能では，指定したキーワードの検索結果順位を時系列で確認する．本機能では，利用者が指定したキーワード毎に検索結果の順位を蓄積し，蓄積した情報を時系列で確認する．

検索結果順位情報は，Web サイトの変更箇所，売上高やアクセス数など，その他の資料と共に時系列で管理し，グラフや表形式で可視化することにより，新たな知識の創造に利用できる．

5. 実験計画

本実験では，検索エンジンから有用なメタ情報を創出できたかを確認する．そのため，まず，検索エンジンを解析し，SEO 対策やサイト管理に必要な情報を創出する．次に，検索エンジンから取得した情報が実際に有用であるかどうかを確認するために，「テックジャム社(http://www.tech-jam.com)」の協力を得て，本システムの実証実験を行った．テックジャム社は，理化学機器のインターネット販売を手がけており，月 13 万人のアクセス数を記録している EC サイトを運営している．実験期間は，2003 年上半期と下半期とした．

実験内容としては，まず，Web リンクマップ，キーワードの有効性の算出と検索結果順位の各種機能を利用して，検索エンジンからメタ情報を取得する．次に，取得したメタ情報が有用であるかを検証するために，SEO 対策を随時実施し，アクセス数の遷移を調査した．

6．実験結果の評価

実験結果の評価では，まず，検索結果から適切な情報を取得し，Web リンクマップを作成する．次に，取得した情報が有用であるかを検証する．

6.1 検索結果からの情報の取得

検索結果からの情報の取得では，検索エンジンの検索結果から，Web リンクマップを作成する．また，作成した Web リンクマップを基に，FOAF のフォーマットに従って Web サイト間の関係をアーク図で表現した．アーク図とは，RDF のステートメントを可視的に表現するための図であり，FOAF の内容を可視化する場合に有用である．テックジャム社のサイトにリンクしている Web サイトを表現したアーク図の一部を図５に示す．

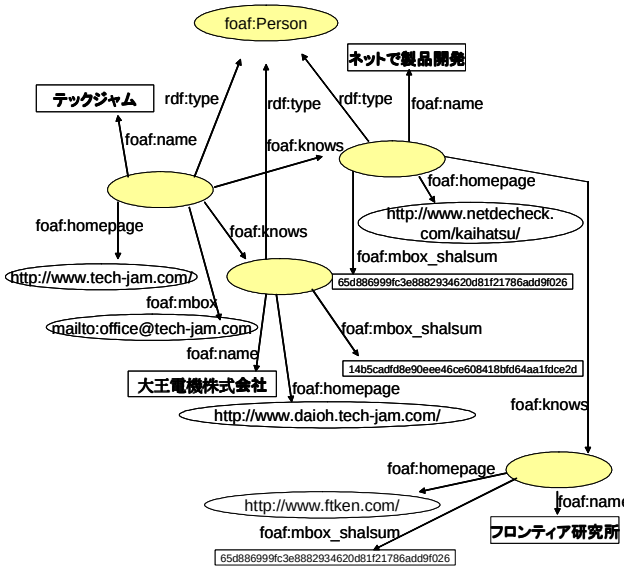

図５　テックジャム社のサイトのアーク図

図５では，テックジャム社は，「ネットで製品開発」と「大王電気社」のサイトとリンクしており，これらの企業

は，テックジャム社と関わりがあることを示している．また，「ネットで製品開発」と「フロンティア研究所」は，「ネットで製品開発」を介して，間接的にテックジャムに関わりがあることを示している．これらの付加情報を基に，ユーザの流れを分析し，顧客を引き入れるための SEO 対策もしくは事業化計画を立案できる．

6.2 検索結果からの情報の有用性の評価

本実験では，Web リンクマップ，検索結果順位分析とキーワードの有効性の算出機能が，実際に有用であるかを検証するために実験システムを利用して SEO 対策を実施した．その後，実験システムの導入前と導入後のアクセス数の遷移を蓄積し，本研究が有用性についての評価を行った．

- 実験システム導入前の 1999 年上半期から 2002 年下半期までのアクセス数の遷移を基に指数近似曲線を算出
- 指数近似曲線の数式を基に 2003 年の上期と下期のアクセス数の予測値を算出
- 指数近似曲線の値を基に算出したアクセス数の予測値と実際のアクセス数を比較し，アクセス数の変化を評価

実験システム導入前の 1999 年上半期から 2002 年下半期までのアクセス数の遷移を基に算出した指数近似曲線と実際のアクセス数のグラフを図６に示す．

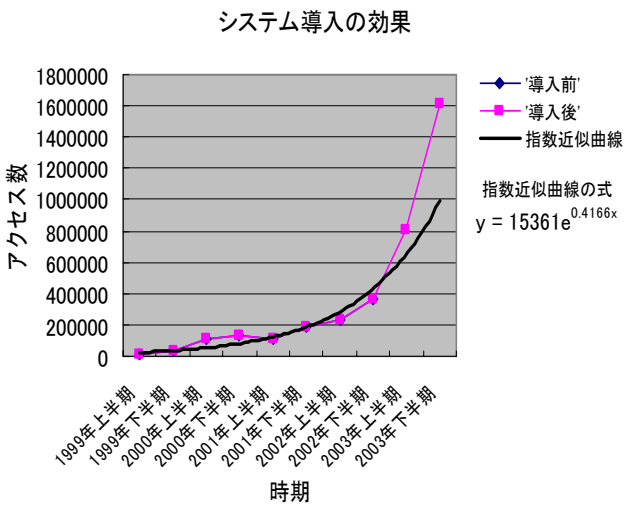

図６　指数近似曲線と実際のアクセス数のグラフ

1999 年上半期から 2002 年下半期までのアクセス数を基に指数近似曲線を出力すると図６のような結果が得られた．また，図６の指数近似曲線の数式を基に 2003 年の上期と下期のアクセス数を予測したところ，表１の様な結果が得られた．

表１　アクセス数を予測値と実測値

時期	予測値	実測値
2003 年上半期	652,948 人	807,045 人
2003 年下半期	990,416 人	1,616,538 人

表１から，実際のアクセス数が指数近似値よりも 2003 年上半期では 154,097 人，2003 年下半期では 686,122 人，上回っているという結果が得られた．このことから，本システムによって，検索エンジンから取得した情報が有用であることが証明できた．

7．おわりに

本研究では，検索エンジンから SEO 対策に有用な情報を創出することを実現した．実証実験を行った結果，検索エンジンから有用な情報を抽出できることを確認した．また，テックジャム社の協力の下，抽出した情報の有用性を確認したところ，アクセス数が，指数近似値の予測値よりも遥かに高い数値を記録した．このことから，本実験システムの導入による効果は大きく，社会的貢献度も高いことが分った．

今後の課題としては，Web リンクマップ機能によって生成された FOAF フォーマットの文書を利用して，Web サイトの信頼度を算出できるようにすることである．本機能が実現することにより，セマンティック Web の根幹となっている RDF の可能性を向上できる．

謝辞

本研究の実証実験を実施するに当り，テックジャム社の迫田博文氏，平尾真一氏，福井貴啓氏からから多大なるご協力を賜り，深く感謝する次第であります．

参考文献

[1] 総務省：平成 15 年度版情報通信白書，pp.2-36，2003.7
[2] 福島俊一： Web サーチエンジンの基本技術と最新動向(下)最新技術，情報管理，科学技術振興事業団，Vol.46，No.7，pp.436-445，2003.10
[3] 関裕司：インターネット検索エンジン 利用者側から見た Google の特徴と使用方法，情報の科学と技術，情報科学技術協会，Vol.54， No.2，pp.90-94，2004.2.
[4] 安藤幸央：人と人をつなぐソーシャルネットワーキング 注目のソーシャルネットワーキングとは?，Software Design，技術評論社，No.162，pp.1-4，2004.4.
[5] World Wide Web Consortium ： RDF/XML Syntax Specification, http://www.w3.org/TR/rdf-syntax-grammar/
[6] 荻野達也：セマンティック Web の最新動向，情報処理学会シンポジウム論文集，情報処理学会，Vol.2003，No.17，pp.9-31，2003.11.
[7] 山名早人，近藤秀和：サーチエンジン Google，情報処理学会誌，情報処理学会，Vol.42，No.8，pp.775-780，2001.8.
[8] 渡辺隆広：検索にガンガンヒットするホームページの作り方，翔泳社，2003.4.