サイト品質管理のためのリンク不整合検出

Bad Link Detection for Web-site Management

河合 英紀♥ 河野 泉♦ 石黒 義英◆ 福島 俊一◆

Hideki KAWAI Izumi KOHNO
Yoshihide ISHIGURO Toshikazu FUKUSHIMA

近年、企業サイトの巨大化・複雑化に従い、リンク不整合の発生が深刻化している。リンク不整合には、デッドリンクなどの物理的不整合と、間違いリンクなどの論理的不整合の2種類がある。物理的不整合は、サーバやブラウザで発生するエラーを検出することにより、従来のサイト管理ツールで検出が可能である。一方、論理的不整合はエラーを発生しないため従来の管理ツールでは検出できない。

そこで本研究では、複数ページ間でのリンク元、リンク先、アンカー文字列の同一性と仲間外れに着目して論理的不整合を効率よく検出する方法を提案する。また、代表的な大規模企業 Web サイトにおけるリンク不整合発生の実態調査を行い、検出された不整合を体系的に分類し、サイト品質管理上のガイドラインを示した。

Recently, the size of websites has grown and website management has become more complex. It also raises the serious problem of link integrity. The referential violation of links falls into two categories: physical bad links and logical bad links.

In this paper, we propose a method of detecting logical bad links, which is based on the identity and difference of a link's source page, target page, and anchor text between multiple pages. We investigated realistic conditions of link integrity on enterprise websites. We also proposed new website management guidelines based on the taxonomy of bad links.

1. はじめに

近年、インターネットの普及に伴って顧客や株主に対する情報発信の窓口としてWebサイトを構築する企業が増えている。企業の顔ともいえる企業Webサイトは、より新鮮で詳細な情報を利用者に提供するために、頻繁に更新を繰り返しながら巨大化している。そのため、サイトの品質を保つことが難しくなってきている。

サイトの品質低下に伴って発生する問題として顕著なのが、デッドリンクや間違いリンクなどのように、本来意図された効果とは異なる結果をもたらしてしまうリンク不整合である。リンク不整合の問題は古くから重要視されており、ジョージア工科大学によるWWWユーザーサーベイ[1]では、Web利用上の重大な問題点として、「デッドリンク」が3位に入っている。また、Pitkowら[2]は、AOLのサーバのログに残ったエラーから、リクエストされたリンクの5～8%がデッドリンクであることを報告している。リンク不整合が多いと利用者の不満は高まり、サイトだけでなくその企業への信頼感も損ねることになりかねない。

リンク不整合は、大きく物理的不整合と、論理的不整合の2種類に分けることができる。物理的不整合とは、デッドリンクのように文書にアクセスした時点で物理的なエラーが発生する不整合である。物理的不整合は、このエラーを検知できれば発見が容易であるため、検出ツールも数多くリリースされている[3,4]。また、検出だけでなく、リンク切れを自動修正する研究[5,6]も数多くなされている。

一方、論理的不整合は、間違った製品情報やサービスへのリンクなど、物理的にはアクセス可能であっても、論理的な誤りを生じている不整合である。これら論理的不整合は、文書にアクセスした時点でエラーは発生しないため、エラー検知に基づく従来の検出ツールでは発見できない。このような論理的不整合の検出方法としてBuchananら[7]は、アンカー文字列とリンク先ページの内容の意味的な整合性を判定するツールを開発しているが、適用範囲は150ページ以下の小規模サイトに止まっている。また、従来はリンク不整合を検出しても対処療法的にそれを修正するだけであり、サイト管理上の問題点について考察されることはあまりなかった。

そこで本研究では、複数ページ間でのリンク元、リンク先、アンカー文字列の同一性と仲間外れに着目して論理的不整合を効率よく検出する方法を提案する。また、代表的な大規模企業Webサイトにおけるリンク不整合発生の実態調査を行い、検出された不整合の実例に基づいたサイト品質管理上のガイドラインを示す。

2. 論理的不整合の種類

人手によるしらみつぶしチェックに基づく予備実験の結果、論理的不整合を次の 3 種類に分類した。

(A)間違いリンク

間違いリンクは、アンカー文字列から期待される内容と、リンク先の文書の内容が異なるリンクである。

間違いリンクの例を図 1 に示す。図 1 では、Page C、D、E からアンカー文字列「Y4100」で、製品 Y4100 の詳細情報である Page A へ正しくリンクが張られている。ところが、Page F からはアンカー文字列「Y4100」で製品 X3100 の詳細情報である Page B へ間違ってリンクが張られている。

(B)リンク元表記の不統一

同一文書を指す複数のリンクのアンカー文字列にゆらぎがあると、利用者が混乱する原因となる。英語・漢字・仮名遣いなどの文字のゆらぎのほかに、アンカー文字列のスペルミスや「新着情報」「お知らせ」のように類似の意味を持つ用語の混乱などもリンク元表記の不統一に含む。

(C)幽霊リンク

幽霊リンクは、HTML では<A>タグでリンクが設定されているにも関わらず、アンカー文字列が設定されていないリンクである。サイト管理者はブラウザで目視確認できないが、検索エンジンで用いられる Web クローラはリンク先のページに容易にアクセスできてしまう。そのため、リンク先に機密情報や古い価格情報などが指定されていると、管理者の気づかないところでそれらの情報が検索エンジンにインデックス

♥ 正会員 NEC インターネットシステム研究所
h-kawai@ab.jp.nec.com
♦ NEC インターネットシステム研究所 {kohno@ay, ishiguro@cw, t-fukushima@cj}.jp.nec.com

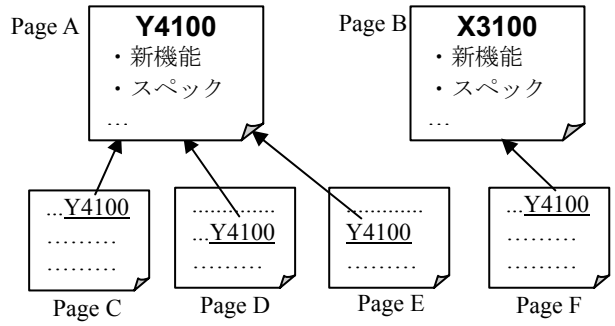

図 1 間違いリンクの例
Fig.1 Example of Misleading Link

されてしまう可能性がある。

3. 論理的不整合の検出ルール

前節で述べた 3 種類の論理的不整合の検出方法として、複数ページ間でのリンク元、リンク先、アンカー文字列の同一性と仲間外れに着目する方法を提案する。

図 1 のような間違いリンクの場合、アンカー文字列が「Y4100」であるリンクをグループ化すると、リンク先は 3 件が文書 A であり、1 件だけが文書 B で仲間外れになる。つまり、(1)アンカー文字列が同一であるにもかかわらずリンク先の文書が仲間外れになるリンクは間違いリンクである可能性があると仮定できる。また、(2)リンク先が同一であるにもかかわらずアンカー文字列が異なるリンク、(3)同一文書内のリンクのうちリンク先文書が同一であるにもかかわらずアンカー文字列が異なるリンク、(4)同一文書内のリンクのうちアンカー文字列が同一であるにもかかわらずリンク先文書が異なるリンク、も間違いリンクの可能性がある。

同様に、リンク元表記の不統一も間違いリンクと同じ(1)～(4)のルールで検出できる。

さらに、 (5)アンカー文字列がヌル文字やスペースのみのリンクは幽霊リンクとして検出できる。

論理的不整合の検出方法として、我々は他にも、個別のリンクについてアンカー文字列とリンク先文書の内容を比較する方法や、文書とリンクの関係の時系列変化を観察する方法なども検討した。しかし、提案方式では、リンクをグループ化することにより、(i) 論理的に正しいリンクと間違っているリンクを明確に区別できる、(ii) 一度のチェックで同じ種類の不整合をまとめて検出できる、という利点がある。従って提案手法では、大規模サイトでも効率よくチェックすることができる。

4. 実験

提案方式を実装した実験システムを構築し、実際の企業 Web サイトにおけるリンク不整合発生の実態調査を行った。診断対象としたサイトは、電気・輸送用機器・食料品・サービスなど、異なる業種における様々な規模の企業 Web サイト 12 件で、実験期間は 2002 年 5 月 17 日～2003 年 9 月 26 日までである。リンク不整合の検出は次の手順で行った。

1) 診断対象のサイトの全文書を Web クローラで取得する。この時、エラーを発生したリンクは物理的不整合として計数した。

2) 全文書に含まれる各リンクについて、リンク元文書、リンク先文書、アンカー文字列をデータベースに格納する。

3) 前節の不整合検出ルール (1) ～ (5) に該当するリンクをデータベースから抽出する。

4) 抽出されたリンクを目視確認し、(a)修正が必須な不整合か、(b)不整合だが些細なので修正は必要としないか、(c)不整合でないかを判定する。例えば、「e-mail」と「E-mail」のような表現の揺れなど、大きな問題にはならないと思われる不整合は(b)に分類した。また、検出ルール(5)で検出されたリンクは自動的に(a)に分類した。

また、手順 1 で物理エラーが発生したリンクを物理的不整合としてカウントした。

論理的不整合検出ルール(1)に該当するリンク情報の例を表 1 に示す。表 1 は、アンカー文字列が「Y4100」であるリンクをグループ化しており、そのうちリンク先が Page A のリンクが 2570 件、リンク先が Page B のリンクが 55 件となっている。Page C は 2570 件のリンクのうち任意に選んだ代表のリンク元文書であり、Page F は 55 件のリンクのうち任意に選んだ代表のリンク元文書である。

手順 4 の目視確認では、まず Page C を開いてアンカー文字列「Y4100」のリンクをクリックする。リンク先として Page A が問題なければ、これら 2570 件のリンクは正しいと判定される。次に、Page F を開いてアンカー文字列「Y4100」のリンクをクリックし、リンク先として Page B が間違っていれば、これら 55 件のリンクを修正が必須な不整合と判定する。

最終的に、(a)と分類されたリンクを論理的不整合として計数した。 なお、検出ルールの包含関係によって重複して不整合とされたリンクは除いてある。

表 1 各企業 Web サイトにおけるリンク不整合
Table 1 Example of Detected Bad Links

リンク数	リンク元	リンク先	アンカー文字列
2570	Page C	Page A (Y4100 製品情報)	Y4100
55	Page F	Page B (X3100 製品情報)	Y4100

5. 結果および考察

5.1 各企業 Web サイトにおけるリンク不整合

実験の結果、実際の企業 Web サイトに驚くほど多数のリンク不整合が存在していることが明らかになった。

表 2 に、各企業 Web サイトで検出されたリンク不整合の件数を示す。表 2 では、ページ数は Web クローラで取得できた Web サイトのページ数であり、この順番で Web サイトを並べてある。また、リンク数は、取得したページから抽出されたリンクの数である。物理的不整合は、前節の手順(1)で検出された物理的不整合の件数、論理的不整合は、前節の手順(4)で (a)修正が必須な不整合と判定されたリンクの数である。

物理的・論理的不整合の検出数はサイトの規模に依存しており、ページ数が 5000 件を超えるサイトでは、物理的・論理的不整合の総和が 1000 件を超えるサイトが 7 サイト中 6 サイトもあった。個別のサイトを見る限りでは、あるサイトでは物理的不整合が多かったり、別のサイトでは論理的不整合が多かったりと偏りが見られる。しかし、12 サイト全体で

表 2 各企業 Web サイトにおけるリンク不整合

Table2 Bad Links in Enterprise Websites

サイト	ページ数	リンク数	物理的不整合	論理的不整合	不整合合計
A	18,389	1,263,562	2,766	2,287	5,053
B	17,900	372,322	494	4,570	5,064
C	16,560	151,195	1,086	3,135	4,221
D	16,393	219,318	330	556	886
E	12,161	221,430	1,511	711	2,222
F	8,430	191,437	7,236	1,277	8,513
G	5,450	49,141	39	1,276	1,315
H	4,072	52,701	358	430	788
I	1,532	9,977	78	8	86
J	1,130	35,796	0	873	873
K	339	11,798	4	21	25
L	331	7,026	1	110	111
合計	102,687	2,585,703	13,903	15,254	29,157

見つかった物理的不整合の合計は 13903 件に対して、論理的不整合の合計は 15254 件であった。従って、物理的不整合だけでなく論理的不整合も同様に重要な問題であるといえる。

今回の調査に要した目視確認の回数は 1 サイト当り平均 737 回で、これにより平均 58027 件のリンクの整合性を判定できた。従って、リンクを 1 件ずつしらみつぶしにチェックするのに比べて検出効率は約 79 倍(=58027/737)であった。これは、4 節の実験手順で示したように、1 度のチェックで複数のリンクをまとめてチェックできるからである。提案方式を使うことによって、大規模サイトにおけるリンクの整合性を効率よくチェックすることができたと言える。

また、12 サイトにおける 1 ページ当りの平均不整合合計は 0.28 件(=29157/102687)であった。これは、3〜4 ページに 1 件の割合で、何らかの不整合が含まれていることを示唆している。

5.2 検出されたリンク不整合の具体例

実際の企業 Web サイトで検出されたリンク不整合の実例として、Web サイト運営上の観点から特に注目すべき 4 種類のケースについて、詳しく説明する。

(A) ページテンプレート中の不整合

ページテンプレートとは、サイト全体を通じてレイアウトの一貫性を保つために使われる Web ページの構成部品である。ページテンプレートは、サイトの内のどのページでも同じ部品が使いまわしされるため、元になるテンプレートが間違っていると、多数のリンク不整合を発生させる原因ともなる。

例えばサイト E では、424 件ものページでフッター部分の「Privacy Policy」へのリンクがデッドリンクになっていた。また、サイト C でもフッター部の複数のリンクが全て間違っているページが 20 件程度見つかっている。パンくずリストやローカルナビゲーション内のリンクで、本来飛ぶべき階層とは異なる階層へリンクする間違いや、異なる製品情報へ間違ってリンクするケースが、サイト E、H、J、K など、複数のサイトで見つかった。また、サイト F の 7236 件の物理的不整合のうち、約 1000 件がページテンプレート中であった。

また、サイト G では、画像とテキストからなるリンクのリストで、テキストの方のリンク先は正しいのに、画像のリンク先がリストの前の項目と同じままであるために間違っているリンクが見つかった。これは、リンクをコピー&ペーストして編集した際に、画像のリンク先を変更するのを忘れたことが原因であると考えられる。ページテンプレートだけでなく、このような部品の使いまわしには注意が必要である。

(B) 用語のあいまいな定義が原因の不整合

2 節(B)リンク元表記の不統一の例でも触れたが、用語の定義があいまいであったり、サイト管理者間でコンセンサスが十分にとれてなかったりすると、リンクの一貫性が保てないために、利用者を混乱に陥れることになる。

サイト E では、グローバルナビゲーションの中に「新着情報」というリンクがあるが、5000 件以上は「新着情報」のページにリンクしているのに対し、169 件が「イベント・セミナー」へのリンクになっていた。さらに、同じグローバルナビゲーション中の「お知らせ」というリンクは、5000 件以上が「イベント・セミナー」へのリンクになっているのに対し、289 件が「プレスリリース」へのリンクになっていた。これらは、「新着情報」「お知らせ」「プレスリリース」「イベント・セミナー」など、新しい情報を提供するような意味を漠然と表すアンカー文字列であるために、一部のサイト管理者が勘違いして間違ったページにリンクしてしまったのだと考えられる。

その他のサイトのケースでは、サイト I では、IR 情報提供のページで、「貸借対照表」へのリンクを「四半期業績レポート」へ間違ってリンクしていた。また、サイト J でも「News」と「Press Release」とを混同しているリンクが見られた。

このような不整合に対しては、類似の情報をまとめてひとつのコンテンツとして扱うか、各用語の意味を細かく定義し、各サイト管理者に徹底させるなどの施策が必要である。

(C) サイトリニューアル時に発生する不整合

企業 Web サイトでは、規模やトラフィックが拡大すると、時々大幅なリニューアルが行われる。サイトリニューアルでは、必ずしもサイト全体のページが一度に全部置き換わるわけではない。むしろ、企業の組織編制の変化に追随してサイトの一部の構成が組み変わったり、特定のコンテンツが独立して別のサーバで運営されるようになったりするような場合が多い。そのため、リニューアルで新しく置き換わるページと、リニューアルでも変化しないページの間で不整合が大量に発生することになる。

最も顕著な例がサイト F であり、7236 件のデッドリンクのうち、80%以上がサイトリニューアル後、旧コンテンツのメンテナンスを放棄しているために発生している。サイト F では、以前はプレスリリースや製品情報などを一つのサーバでまとめて運営していた。ところがある時、製品情報を新規サーバに移して独立に運営するようになったが、この時、旧サーバに残ったコンテンツ内のリンクを修正しなかったため、大量のデッドリンクが発生することになってしまった。

また、サイト E でも、サイトリニューアル直後に新旧タイトルのアンカー文字列を持つリンクが混在する不整合が観察された。本ツールでは、サイトリニューアル時の大規模なリンク関係の変更もチェックすることができる。

(D) デッドリンクの隠蔽

表 2 では、サイト J における物理的不整合は 0 件であった。

確かに、ページ収集中に「404 Not Found」に相当するエラーは1件も出ていなかった。しかし、論理的不整合のチェック時に、アンカー文字列で示された内容が含まれないのにトップページにリダイレクトされるリンクが数件見つかった。試しに、サイトJ内ででたらめなURLを指定してみてもエラーにはならず、トップページへリダイレクトされた。つまり、このサイトでは、問い合わせのあったURLがない場合でもエラーを返さず、代わりにトップページへリダイレクトするようにHTTPサーバを設定していたことになる。クローラのログを確認したところ、リンク先が存在しないためにリダイレクトされているリンクは58件であった。

この「デッドリンクの隠蔽」はリンク不整合対策としては、次の2つの点で良くない方法である。一つ目は、アクセスログなどを見ない限り、サイト管理者自身もデッドリンクの存在に気づかないために、デッドリンクが放置されてしまうこと。もう一つは、デッドリンクと分からないサイト閲覧者が、リダイレクトされたトップページで、本来はないはずの情報を探すために時間を費やさなければならないからである。本ツールでは、このようなケースも検出可能である。

6. リンク不整合に基づくサイト管理指針

実験結果から明らかになったサイトの品質管理に関する新しいガイドラインを提言する。

(1) リンクの物理的不整合だけでなく、論理的不整合にも気をつけるべきである。その理由は、5.1.節で示したように、物理的不整合と同程度の割合で論理的不整合も発生しているからである。

(2) メンテナンスが優れているサイトとは、1ページ当りのリンク不整合 0.28 件より少ないサイトである。その理由は、5.1 節で示したように、平均的なサイトのリンク不整合の発生割合は1ページ当り0.28件だからである。

(3) コンテンツマネジメントシステムを使っていても、リンク不整合には気をつける必要がある。その理由は、5.2節(A)で示したように、テンプレートが間違っていると、リンク不整合を多数発生させる原因となるからである。

(4) サイト内で使われる用語について、管理者間で明確な定義がなされているか確認すべきである。その理由は、5.2節(B)に示したように、あいまいな用語がリンクの論理的不整合発生の原因となっているからである。

(5) サイトをリニューアルする場合、リニューアル対象でないページでも、リンクを改めてメンテナンスすべきである。その理由は、5.2節(C)に示したようにサイトリニューアル後に、リニューアル対象でないページのリンクが不整合の原因になることが多いからである。

(6) リダイレクトを使ってデッドリンクを隠蔽すべきでない。その理由は、5.2節(D)で触れたように、サイト管理者自身がデッドリンクのメンテナンスを放棄した分のコストを、サイト閲覧者がかぶらなければならないからである。

7. おわりに

本論文では、Webサイトにおける論理的不整合の効率のよい検出方法を提案し、実際の企業Webサイトにおいて多数のリンク不整合が存在する実態を示した。また、実際に検出された多くのリンク不整合の実例を紹介し、サイト品質向上のための新しいガイドラインを示した。

今後は、検出精度の向上と自動化を図るとともに、不整合検出結果から原因推定・サイト改善提案へ結びつける分析手法を検討していく。

[文献]

[1] Graphics, Visualization, & Usability Center, GVU's Tenth WWW User Survey, Question 11, "Problems Using the Web" ,1998. http://www.gvu.gatech.edu/user_surveys/survey-1998-10/graphs/use/q11.htm

[2] J. Pitkow. Web Characterization Activity Answers to the W3C HTTP-NGs Protocol Design Group's Questions. World Wide Web Consortium, 1998. http://www.w3.org/WCA/Reports/1998-01-PDG-answers.htm

[3] Watchfire, http://www.watchfire.com/default.aspx

[4] LinkScan, http://www.elsop.com/linkscan/

[5] Helen Ashman, Electronic Document Addressing - Dealing With Change. *ACM Computing Surveys*, 32(3), p.201, September 2000.

[6] Seung-Taek Park, David Pennock, Lee Giles, Robert Krovetz, Analysis of Lexical Signatures for Finding Lost or Related Documents , *Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, p. 11, 2002.

[7] G. Buchanan, G. Marsden, H. Thimbleby, Meaningful Link Verification for Management and Maintenance of Web Sites, *In Proceedings of the 8th International World Wide Web Conference (WWW8)*, Toronto, May 1999.

河合 英紀 Hideki KAWAI

1998 年慶應義塾大学理工学研究科修士課程卒業。同年日本電気株式会社入社、NEC インターネットシステム研究所に所属、現在に至る。情報の検索および構造化の研究・開発に従事。情報処理学会会員。日本データベース学会会員。

河野 泉 Izumi KOHNO

1991 年大阪大学基礎工学研究科修士課程卒業。同年日本電気株式会社入社。現在、NEC インターネットシステム研究所に所属。ヒューマンインタフェースの研究・開発に従事。情報処理学会会員。日本デザイン学会会員。

石黒 義英 Yoshihide ISHIGURO

1990年京都大学大学院工学研究科電気工学第二専攻修了。同年、NEC入社。グループウェア、エージェント、情報検索等に関わる研究開発に従事。1998-1999年ジョージア工科大学客員研究員。現在、NECインターネットシステム研究所主任研究員。ACM、情報処理学会、各会員。

福島 俊一 Toshikazu Fukushima

1982年東京大学理学部物理学科卒業、NEC入社。現在、同社インターネットシステム研究所ユビキタスインテリジェンステクノロジーグループ研究部長。工学博士。情報処理学会第6回坂井記念特別賞・平成4年度論文賞、第51回オーム技術賞ほか受賞。情報処理学会、人工知能学会、ACMほか会員。