リンク不整合検出によるWebサイト診断－不整合結果の多面的分析

河野 泉、河合 英紀、石黒 義英、福島 俊一
NEC インターネットシステム研究所

１．はじめに

近年、企業の Web サイトは数万ページ規模のものも多く、内部のリンク総数は数十万に及んでいる。その中で、本来意図したページとは違うページへ遷移してしまう「リンク不整合」は品質上問題である。我々は、デッドリンクのようにアクセス先のページが無くなった「物理的不整合」以外に、間違ったページへリンクが貼られた「論理的不整合」を効率的に検出する手法を提案し、実際のサイトに多くの不整合が存在することを示した[1]。これらの不整合は数も多くサイトの頻繁な更新頻度に合わせて修正するためには、結果の分析が重要である。本稿では、不整合結果の多面的分析方法を提案し、当社サイトの品質改善へ適用した結果を報告する。

２．リンク不整合検出の概要

我々は、文献[1]で提案した不整合検出ルールを実装したリンク不整合検出システム Simprove を開発してきた。リンク不整合検出は、1)クローラによる Web ページ収集、2)リンクの属性情報抽出、3)ルールによる不整合検出、4)検出結果の不整合レベル入力の手順で行われる。

Simprove を使って企業、自治体サイトのリンク不整合実態を検査したところ、表１のように多くの物理的不整合と論理的不整合が存在していた。ページ総数はクローラで収集したページ数、リンク総数は収集したページから抽出したリンク数で、リンク総数が約 10 万以上の大規模サイトの結果を示している。リンク総数に対する不整合総数の割合である不整合比率を見ると、この規模のサイトには平均して 1%程度の不整合が存在していることがわかる。

Website Diagnosis Based on Bad LinkDetection
- Multi-Aspect Analysis of Bad Links
Izumi Kohno, Hideki Kawai, Yoshihide Ishiguro,Toshikazu Fukushima
Internet System Research Labs., NEC Corp.
E-mail:kohno@ay.jp.nec.com

３．サイト品質改善のための多面的分析

リンク不整合はサイト訪問者の目的達成を阻み、商品購入を中断させたり企業イメージの低下を引き起こすこともあるため、迅速な修正が必要である。しかし、大規模 Web サイトは、複数部門による管理、頻繁な更新、限られた労力という制約下で運用されており、多数のリンク不整合を修正するのは容易ではない。従来のリンクチェックやサイトマネジメント製品[2]は、検査が主目的で、修正のための機能はあまり無い。我々は、リンク不整合の修正には適切な修正方針をたて、修正方針に沿った作業を支援する結果の分析が重要だと考える。以下に我々の提案する、不整合の修正方針、分析方法、Simprove へ実装した分析機能について説明する。

1)不整合総数の効率的削減

不整合総数を少ない労力で短期間のうちに減らすことは最も重要である。そのためには、不整合数を種類別に集計し、数の多い問題から解決していく事が有効である。Simprove では、同じアンカー文字列と同じリンク先をもったページがグループ化されて、１種類の不整合として検出される。グループ内のページ数が多い不整合は、そのサイトで多く発生している問題である。効率的な修正のため、不整合数別（グループ内のページ数別）の分析機能を実装した。

2)重要な不整合の優先的修正

限られた労力では、数多くの不整合のうち重要なページにある不整合から優先的に修正していくことも重要な方針である。重要なページとは、アクセス数の多いページや、トップページからの階層が浅いページなどが考えられる。これらのページにある不整合は、訪問者の目につきやすく悪い印象を与えやすい。修正優先順位を決めるため、不整合のページビュー別分析機能、階層別分析機能を実装した。

3)原因特定による再発防止

不整合を個別に修正するだけではなく、根本的な原因を特定し再発を防止することは重要で

ある。そのためには、発生箇所を分析し管理部門を特定することや、不整合の修正状況を分析し管理体制を把握することが有効である。原因特定のため、不整合のディレクトリ別分析機能、時間変化分析機能を実装した。

４．当社サイトへの適用

４．１ 適用結果

当社の社外向け２サイトについて、月１回の定期的なリンク不整合検査と多面的分析を行った。またサイトリニューアル直後にも検査/分析を実施した。その結果、表２のように、当社サイトは不整合比率 0.1 ～ 0.2%台という、同規模のサイトよりかなり少ない状態を維持できた。リニューアル直後に一時的に増加したリンク不整合を素早く修正することも可能であった。

４．２ 考察

不整合結果の多面的分析により、どのような問題が発見され、改善に役だったかを述べる。

1)不整合数による分析

不整合数別分析により、Copyright や PrivacyPolicy というフッタに不整合(174 件)が多く発生するケースが見つかった。ヘッダ、フッタ、メニューなどはサイトの共通部品のため、多くの不整合を引き起こす場合がある。リニューアル直後の分析では、リンク先のページタイトルがリニューアルで変更されたが、アンカー文字列がリニューアル前のタイトルのままという不整合(548 件)が多く見つかった。これらの問題は１つの原因で多くの不整合が発生しており、修正しやすく効率よく改善できた。

2)ページビューによる分析

ページビュー別分析により、アクセス数の多いページにもリンク不整合が多く発生するケースが見つかった（図１）。これは、アクセス数の多いページは更新頻度が高いために起こったと考えられる。これらの不整合は、修正優先度の高い不整合としてサイト管理者に通知した。

3)ディレクトリによる分析

ディレクトリ別分析により、１つのディレクトリに全体の 56%の不整合が集中するケースが見つかった。サイト管理者に問題を指摘したところ、このディレクトリは管理部門が別になっており、サイト全体に行っているはずのデッドリンク（物理的不整合）のチェックから漏れていたため、不整合が多く発生していることがわかった。ディレクトリ別分析により、管理体制に関する原因特定ができた。

５．おわりに

リンク不整合をページビュー別やディレクトリ別など多面的に分析することにより、修正の優先度や原因が特定でき、サイトの品質改善に役立つことを示した。定期的な診断により、当社サイトは他社より不整合比率を削減できた。

参考文献

[1] 河合英紀他, リンク不整合検出による Web サイト品質評価, LD-005, FIT2003, 2003 年 9 月.

[2] http://www.elsop.com/linkscan/

表1 他社サイトのリンク不整合結果　　調査時期 2003年7月～10月

他社サイト	ページ総数	リンク総数	物理的不整合	論理的不整合	不整合総数	不整合比率
企業A	22,940	737,293	6,447	2,059	8,506	1.15%
企業B	19,019	165,245	805	672	1,477	0.89%
企業C	4,741	139,384	653	53	706	0.51%
企業D	3,853	107,888	422	227	649	0.60%
自治体A	10,753	221,197	2,272	1,160	3,432	1.55%
自治体B	12,302	96,067	980	573	1,553	1.62%

表2 当社サイトのリンク不整合結果　　調査時期 2003年7月～10月

当社サイト	ページ総数	リンク総数	物理的不整合	論理的不整合	不整合総数	不整合比率
A (7月)	16,743	232,685	320	197	517	0.24%
A (10月)	15,974	212,005	349	222	571	0.25%
B (7月)	16,215	335,932	526	399	925	0.28%
B (10月)	20,022	495,310	341	221	566	0.11%
B (リニューアル直後)	17,059	340,329	585	2,938	3,523	1.04%

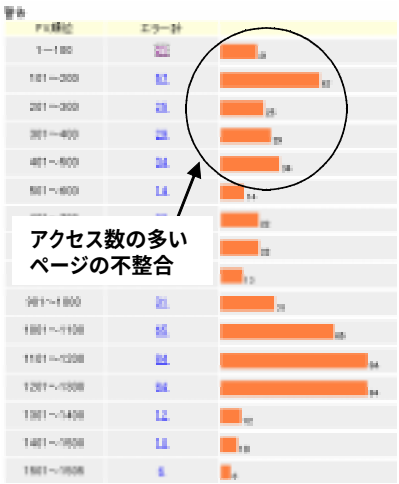

図1 ページビュー別リンク不整合結果