リンク不整合検出による Web サイト診断 ー論理的不整合の自動判定

河合英紀　河野泉　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 Web サイトの大規模化･複雑化に伴い、リンク不整合の問題が深刻化している。リンク不整合には、デッドリンクのように文書にアクセスした時点でエラーが発生する不整合(物理的不整合)と、明示的なエラーは発生しないが論理的な誤りを生じている不整合（論理的不整合）の 2 種類がある。

筆者らは、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した検出ルールを提案し、実際の企業 Web サイトでは、物理的不整合に加え論理的不整合も多数存在することを示してきた[1]。そのなかで、単純な検出ルールでは、論理的不整合の候補として検出されたリンクが必ずしも不整合でない場合もあるため、候補を人手でチェックする作業の効率向上が望まれていた。そこで本稿では、論理的不整合の候補に対して、機械学習の一手法である Support Vector Machine(SVM)[2]を用いてそのリンクの整合性を自動判定する方法を提案する。

2. 論理的不整合検出方法とその問題点

図 1 に､論理的不整合の例を示す｡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。ここで、検出ルール(1)「アンカー文字列が同一のリンクのグループ内で、リンク先文書が異なるサブグループ」を適用すると、Page C～F に含まれるアンカー文字列が「Y4100」の 4 つのリンクのグループ内で、Page A をリンク先とする 3 つのリンクと、Page B をリンク先とする 1 つのリンクが異なるサブグループとなり、論理的不整合の候補として検出される。これら 2 つのサブグループに対して、それぞれアンカー文字列とリンク先の整合性を人手でチェックすることにより、Page F からのリンクが論理的不整合であると判断される。文献[1]では検出ルール(1)の他に 4 つの検出ルールを用いた。

一方､上記の検出ルールは非常に単純であるため、論理的不整合の候補として検出されたリンクが必ずしも論理的不整合でない場合もある。例えば、アンカー文字列が「戻る」「次へ」などのナビゲーション用リンクは、同じアンカー文字列で異なる文書にリンクされる傾向があるため､検出ルール(1)で誤検出されてしまう。誤検出を削減するためには、検出ルールをさらに詳細化する方法が考えられる｡しかし、実際の論理的不整合のケースは多様であり、全てのルールを網羅的に記述するのは困難である。そこで本稿では、機械学習の一手法である SVM 用いてリンクの整合性を自動判定することにした。

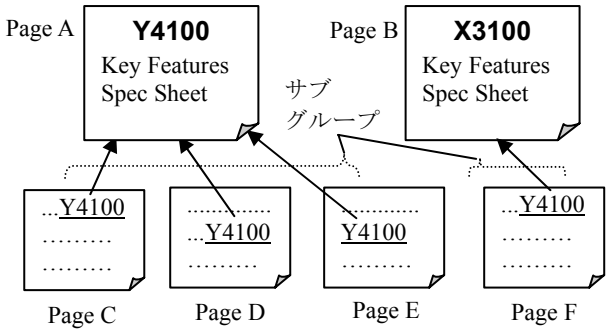

図 1 論理的不整合の例

3. SVM による論理的不整合の自動判定

SVM[2]は「マージン最大化」を基準として分離平面を決定する 2 値分類器であり、他の機械学習の手法に比べて優れた認識性能を発揮できるとして、近年注目を集めている。

リンクの整合性判定は、診断対象とするリンクの素性を入力とし、そのリンクが不整合か否かを出力とする 2 値分類問題と考えることができる。本稿では、検出ルールに該当したリンクのサブグループを診断対象とし、リンクの素性として次の 2 種類を用いることとした。

(1) リンクグループに関する素性

図 1 のような論理的不整合が存在する場合、検出ルール(1)でグループ化するとサブグループに含まれるリンク数に偏りが出る。この時、少数派のサブグループに属するリンクが不整合の可能性が高いと考えられる。そこで、(1a)グループ化に使った検出ルール、(1b)グループ内における各サブグループのリンク数の分散、(1c)リンクグループのリンク数に対するサブグループのリンク数の比を素性とした。

(2) リンク関係に関する素性

間違いリンクなどの論理的不整合では、アンカー文字列とリンク先文書で使われている単語が一貫し

Website Diagnosis Based on Bad Link Detection - Automatic Detection for Logical Bad Links
Hideki KAWAI, Izumi KOHNO, Yoshihide ISHIGURO, Toshikazu FUKUSHIMA
Internet Systems Research Laboratories, NEC Corp.
E-mail:h-kawai@ab.jp.nec.com

ない傾向があると考えられる。そこで、各リンクサブグループについて、アンカー文字列中の単語がリンク先文書において、(2a)タイトルに含まれる割合、(2b)本文に含まれる割合、(2c)H タグや FONT タグ等によって強調表示されている割合を素性とした。

リンクの整合性を判定するための素性としては、他にもリンク先 URL に含まれるディレクトリ名やアンカー文字列に含まれる単語なども考えられる。しかし、予備実験では上記 2 種類の素性に限定した方が、アンカー文字列や URL を素性として追加するよりも優れた判別精度を示した。

また、学習データとしては、(A)診断対象データと同じサイトの過去のチェック済みデータ、(B)診断対象データとは異なるサイトの過去のチェック済みデータ、(C)診断対象データの一部を人手でチェックしたデータ、などの利用が考えられる。本稿では、同じサイトを定期的にチェックする運用形態を想定し、(A)診断対象データと同じサイトの過去のチェック済みデータを学習データとして用いることにした。

4. 実験

実際に 2 件の企業 Web サイトを対象として不整合チェック作業の効率改善実験を行った。サイト A に関しては 2003 年 8 月 15 日に不整合チェックを行った結果を学習データとし、2003 年 9 月 15 日に Web クローラで収集したデータを診断対象データとした。サイト B も同様に 2003 年 8 月 28 日のチェック済みデータを学習データ、2003 年 9 月 26 日のデータを診断対象データとした。SVM プログラムには、TinySVM[1]を用いた。

5. 結果および考察

表 1 に検出ルール単独と SVM を用いた場合の判定精度を示す。検出ルール単独で見つかった不整合総数に対し、SVM が検出できた不整合の割合（再現率）は 50%台であった。しかし、不整合の候補として検出され実際に不整合だった割合（適合率）は検出ルール単独では 20%台だが SVM では 80%台に向上した。

図 2 および図 3 にそれぞれ、サイト A、B において作業時間当たりに検出された不整合数を示す。図 2 の 7 月 15 日、8 月 15 日、および図 3 の 8 月 28 日は SVM を用いずに検出ルールに該当したリンクサブグループを先頭から順にチェックした。一方、図 2 の 9 月 15 日と図 3 の 9 月 26 日は SVM を用いて、分離平面からの距離が大きいサブグループから順に検出ルールに該当したリンクを全てチェックした。いずれの場合においても、SVM を使わない場合は作業時間に従って不整合数はほぼ線形に増加したのに対し、SVM を使った場合は、作業の初期段階で多くの不整合が見つかり、やがて飽和する傾向が見られた。これは、SVM の適合率が高いために、検出ルール単独の場合に比べて短時間で多くの不整合を検出できていることを示唆している。不整合総数の 90%を発見するまでにかかる時間を比べると、サイト A では 8 月 15 日に約 600 分かかったのに対して、9 月 15 日には約 300 分とほぼ半分に短縮でき、サイト B でも約 6 割に短縮できた。

[1] http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

表 1 提案手法の精度評価

サイト	検出ルール	SVM	
	適合率	適合率	再現率
A	22% (153/684)	87% (83/95)	54% (83/153)
B	26% (229/881)	80% (125/156)	55% (125/229)

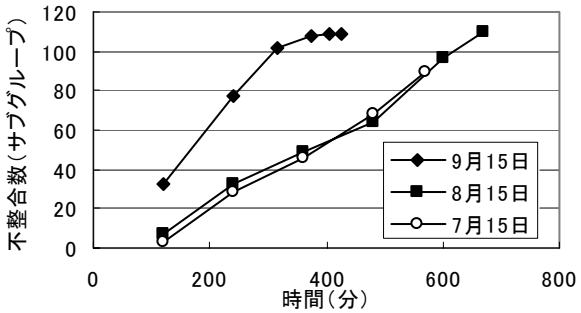

図 2 サイト A における作業効率

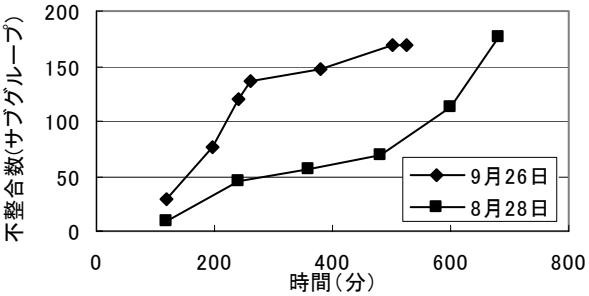

図 3 サイト B における作業効率

6. おわりに

SVM を用いた論理的不整合の自動判定法を提案し、実際の企業 Web サイト上でその有効性を検証した。その結果、再現率は 50%台に低下したが、適合率は 80%台に向上し、不整合発見の時間をほぼ半分に短縮することができた。提案手法は、Web サイトをくまなく診断する用途には向かないが、短時間で効率よくサイトの問題点を発見する用途に適している。

今後は、自動判定の再現率の向上を目指すとともに、他の運用形態も考慮した学習データの選び方について検討していきたい。

参考文献

[1] 河合英紀, 河野泉, 石黒義英, 福島俊一, リンク不整合検出による Web サイト品質評価, FIT2003, LD-005, 2003.

[2] Vladmir N. Vapnik, Statistical Learning Theory, A Wiley-Interscience Publication, 1998.