SD-1-5 Webページの時間順序付け(SD-1. WebとXML ： 新たな展開)

三浦, 孝夫 / 森, 正輝

(出版者 / Publisher)
一般社団法人電子情報通信学会
(雑誌名 / Journal or Publication Title)
電子情報通信学会総合大会講演論文集
(号 / Number)
1
(開始ページ / Start Page)
S-8
(発行年 / Year)
2004-03-08

SD-1-5

Webページの時間順序付け

森　正輝[1]　　三浦孝夫[1]

[1] 法政大学工学部電気電子工学科

1　動機と解決法

現在インターネット上で利用できる検索エンジンの検索サービスには、Directory 検索、Web 検索、Images 検索、Groups 検索、News 検索などがある。News 検索以外の検索では、時間によって内容の変化するキーワード（例：スポーツ・政治・経済）に関しての知識をユーザが持っていなければ Web ページの内容が示す時間をページを拾い読みすることで見つけるしかなく、大量の Web ページの中から知りたい情報を見つけるには多くの時間を費やしてしまう。ニュース検索は、Web ページの文章の内容に関する時間（有効時間）に従った検索結果の表示ができるが、ニュース記事の Web ページを対象にしたものであり、Web ページ全てを対象にしたものではない。他に Web ページに関する時間として、Web ページが更新された時間（トランザクション時間）がある。

本論文では、有効時間に従って Web ページを並べ替える方法を提案する。このため、Web 検索エンジンの検索結果を、各ページの有効時間に従って並べ替える。このことで、これまで自らがページを拾い読みする状況から、その負担が軽減され可読性が向上する。また同時に、トランザクション時間による Web ページの並べ替えでは可読性が向上するわけではないことを示す。

2　並べ替え手順

Web ページの内容の示す時間（有効時間）は以下の手順で決定する。最初に、対象となる Web ページの URL の中に含まれる時間情報を取り出し、その時間情報を Web ページの有効時間とする。この場合、1 つの Web ページが 1 つの有効時間を持つことになる。次に、URL に時間情報を含まない Web ページ中の文章の直前にある日付を取り出し有効時間とする。ここで言う文章とは、単語の後にピリオドその後に”＜”の記号で終わるもの（例：reports. ＜）である。ただし、文章がリンクになっているものは含まない。この場合、1 つの Web ページが複数の有効時間を持つことを許す。対象となる日付の表記は月日年（例：December 9, 2003）とする。

トランザクション時間は各 Web ページの受信ヘッダファイルの「Last − Modified」の値を用いる。

3　実験

実験の対象となる Web ページは、google の Web 検索で英語の Web ページに限定し検索キーワードにヒットした Web ページの TOP100（Weblog は除く）を用い、検索のキーワードは「spirit mars」「iraq war」「kazuo matsui」「nba lakers 2003」「saddam hussein」「bush iraq」「japan earthquake」「japan koizumi」「north korea」を用い 2 章で述べた方法により有効時間とトランザクション時間を取り出し、有効時間による Web ページの並べ替えを行う。

Web ページから取り出した有効時間の整合性を、URL から抜き取った場合、時間情報がが Web ページの有効時間であるかと、Web ページの文書から取り出した場合、取り出した有効時間が Web ページの文章に対して正しい有効時間であったかで判断する。

キーワード	有効時間数	整合性	対象ページ数
spirit mars	69	64/69	100
iraq war	47	46/47	92
kazuo matsui	44	38/44	80
nba lakers 2003	38	35/38	83
saddam hussein	70	66/70	84
bush iraq	69	63/69	94
japan earthquake	40	36/40	98
japan koizumi	60	58/60	90
north korea	66	63/66	95

図 1: 実験結果

	トランザクション時間	有効時間	整合性	アドレス
15	2004/01/09/10:57:22	2004/01/02/0000	○	http://www.xs4all.nl/~carlko
16	0	2004/01/03/0000	○	http://www.space.com/miss
17	0	2004/01/03/0000	○	http://solarsystem.nasa.gov/
18	2004/01/04/09:29:48	2004/01/03/0000	○	http://www.spacedaily.com/
19	2004/01/05/16:59:19	2004/01/03/0000	○	http://www.californiaspacea
20	0	2004/01/03/0000	○	http://www.ctv.ca/servlet/A
21	2004/01/09/10:57:22	2004/01/03/0000	○	http://www.xs4all.nl/~carlko

図 2: 有効時間とトランザクション時間

4　考察

図 1 より整合性は全てのキーワードにおいて、取り出した有効時間が 8 割以上とれている。整合性が取れなかったものは URL から有効時間を取り出したものでページが既に存在しない場合が多く見られ、今回、有効時間を取り出すために用いた方法は機能を果たせた。それにより、時間によって内容の変化するキーワードに対して知識のないユーザの Web ページを拾い読みする負担を軽減でき、可読性を向上させることができた。図 2 はキーワード「spirit mars」の有効時間の古い順に 15 位から 21 位までの有効時間、トランザクション時間、整合性とそのアドレスである。受信ヘッダから「Last-Modified」の値が取り出せなかったものは、トランザクション時間の値を 0 とした。値の取れたトランザクション時間と有効時間を比較するといずれも、有効時間とトランザクション時間では時間が違いトランザクション時間で並び替えた場合と有効時間で並べ替えた場合では順位が変わる。有効時間の整合性が取れているのでトランザクション時間による Web ページの並べ替えでは、可読性は向上しない。

5　結び

本論文では、有効時間に従って Web ページを並べ替えることで Web ページの可読性が向上することを、有効時間を URL 又は文章を含む Web ページから取り出し並び替えを行うことで証明した。今後の課題としては、文章内容および非文章ページから相対順序を抽出すること等があげられる。。

参考文献

[1] 特集セマンティック Web
www.ipsj.or.jp/members/Magazine/Jpn/4307/
情報処理学会学会誌