Internet Killer KILLER

川副　博
kawazoe@trl.ibm.co.jp
(株) 日本アイ・ビー・エム　東京基礎研究所

インターネットの使用者や興味を持つ人の数を増やした理由の一つに World-Wide-Web(以下 WWW と呼ぶ) システム [1] がある。WWW システムは インターネット殺し (Internet Killer) と呼ばれることがある。これは WWW システムサーバ、クライアント間を巨大なトラフィックが流れるからである。画像、音声などの大きくなデータを、クライアントの GUI でのマウスの一操作で転送できる。マルチメディアデータを HTML で容易に記述できるので、WWW システムは情報発信手段として使われることが多い。一方、情報の受け側では発信側で用意したハイパーテキストのリンクを辿る以外にハイパーテキストの全貌を知る手段はない。ハイパーテキストの概要を自動作成し、この概要を使用者がみて、ハイパーテキスト内に必要とする情報があるかどうかを判断させる方法について述べる。

1　背景

インターネットの使用者や興味を持つ人の数を増やした理由の一つに World-Wide-Web(以下 WWW と呼ぶ) システム [1] がある。WWW はサーバ・クライアント型のシステムであり、クライアントは通常、GUI のマルチメディアのハイパーテキスト用ビュウアーである。ハイパーテキストデータはサーバ上に置かれる。　ハイパーテキストのリンクは別のサーバ上のデータへも張ることができる。　このハイパーテキストの記述言語を HTML と言う。WWW システムは インターネット殺し (Internet Killer) と呼ばれることがある。これは WWW システムサーバ、クライアント間を巨大なトラフィックが流れるからである。画像、音声などの大きくなデータを、クライアントの GUI でのマウスの一操作で転送できる。

マルチメディアデータを HTML で容易に記述できるので、WWW システムは情報発信手段として使われることが多い。一方、情報の受け側では発信側で用意したハイパーテキストのリンクを辿る以外にハイパーテキストの全貌を知る手段はない。本研究は WWW システムにおいて情報の受け側で情報の選択方法について提案する。

Internet Killer KILLER
Hiroshi KAWAZOE
IBM Research, Tokyo Research Laboratory

2　目的

本研究の目的はハイパーテキストの中に使用者の欲する情報があるかどうかを知ることを助けることである。ハイパーテキストではリンクの先の情報に関して、そのリンクの先の情報がリンクを辿る前にわかっている場合のみリンク先の情報が使用者が望むものかどうか判断できる。　このためにはハイパーテキストのデータを注意深く作成しなければならない。　ハイパーテキスト内でキーワード検索を行う場合には必ずリンクを辿る必要がある。ハイパーテキストのデータのリンクはネットワークなので、ハイパーテキストの全体を眺めるにはループの検出をしながらリンクを辿る操作を繰り返す必要がある。WWW システムではハイパーテキストデータはネットワークを介して転送される。WWW システムサーバ・クライアントがネットワーク的に近ければ転送時間は気にならないがネットワーク的に遠ければ転送時間は有意なものとなってくる。従って、WWW システム上のハイパーテキストの全体を見るにはリンクを辿る操作、ループの検出が必要であり、リンクを辿る毎の待ち時間がかかるので人間がやるには面倒である。ここではハイパーテキストに自分の欲しい情報があるかどうかを判断するのに必要な情報を自動的に (操作なしで) 作成することを目指している。

3　方法

ハイパーテキストの概要を自動作成し、この概要を使用者がみて、ハイパーテキスト内に必要と

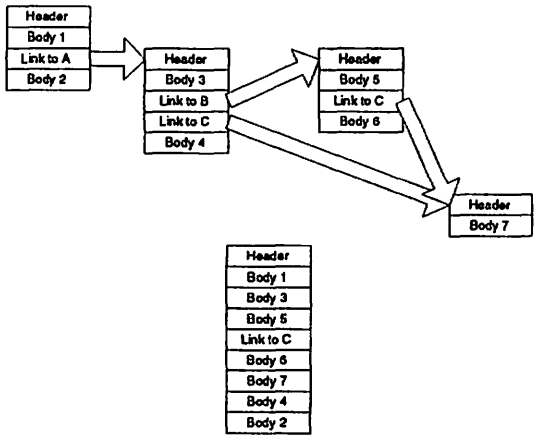

図 1: ハイパーテキストのリンクをハイパーテキストで置き換え概要を作る例

する情報があるかどうかを判断させる。以下では、概要とは何かとどこで概要を自動作成するのかについて述べる。

3.1 概要とは

概要とは次のものを指す。

- 内容に関連のある範囲でハイパーテキストのリンクを、リンクを辿った先のハイパーテキストで置き換えたもの (図 1) から
- 人間がそのハイパーテキスト内に望みの情報があるかどうかを判断するのに不要な情報を落とし、
- 表現の密度を高めた

ものをいう。

リンクを辿る範囲 ハイパーテキストはネットワークなのでループがある。リンクをリンクを辿ったさきのハイパーテキストに置き換えるときに、一度置き換えたリンクは置き換えないで残して置く。

表現の密度 ハイパーテキストのデータ内にはクライアントで表示した時の表現情報 (文字の大きさ、行間、文字種) の指定がふくまれている。表現の密度を高めるとはクライアントの同じサイズの画面により多くのハイパーテキストを表示できるにようにすることを言う。

3.2 概要をどこで作るのか

WWW システムのハイパーテキストは インターネット上に散らばっている。サーバ上のハイパーテキストは先頭として参照されることもあれば別のハイパーテキストの一部として参照されることもある。

概要を作成する場所としては次の 3 点がある。

- 各サーバ
- クライアント
- 概要作成専用サーバ

各サーバで概要を作る場合というのはクライアントからハイパーテキストの先頭があるサーバに対してそのハイパーテキストの概要を作るように依頼する。 依頼を受けたサーバはそのハイパーテキストから順にリンクを辿りリンクをリンク先で置き換えていく。他のサーバ上のデータを指しているリンクにはそのサーバにそのリンクから先の概要作成を依頼する。 この方法ではハイパーテキストのループの検出のためにサーバから他のサーバへ概要作成を依頼する際にループ検出用のデータを渡す必要がある。 そのため、HTTP に変更が必要であり、全サーバががこの変更に対応する必要がある。

クライアントで概要を作る場合は使用者が指定した時にリンクを辿る機能を自動的に実行しリンクを辿る。 この方法ではクライアントに変更が必要である。

概要作成専用サーバの場合はクライアントから専用サーバに概要を作って欲しいハイパーテキストの先頭ノードを指定する。専用サーバはこのノードからリンクを辿りできた結果をクライアントに返す。 この方法では HTTP への変更は必要でなく、サーバ、クライアントとも既存のものが使える。

サーバ、クライアントともに既存のものが使えるので概要作成専用サーバ方式を採用した。

4 現状

3で述べたように概要をつくるためには指定されたハイパーテキストのリンクを開始点とし次の動作を行う。

- 内容に関連のある範囲でリンクを辿り、
- 人間がそのハイパーテキストに必要な情報があるかどうかを判断するのに不要な情報を落とし、
- 表現の密度を高めた

ここでは現在でのそれぞれ「内容に関連のある範囲」と「人間がそのハイパーテキストに必要な情報があるかどうかを判断するのに不要な情報」と「表現の密度」をどのように判定/実装しているかについて述べる。

4.1 内容に関連のある範囲

WWW システムのハイパーテキストのリンクには他のサーバ上のハイパーテキストを指すリンクと同じサーバ上のハイパーテキストを指すリンクとの 2 種類ある。先頭のハイパーテキストより広さ優先でリンクを探し、そのリンクのなかでも自分のサーバ上のデータを指すリンクを優先してリンクとリンク先のデータを置き換える。 この置き換えを使用者が指定した数を限度として行う。

4.2 人間がハイパーテキストに必要な情報があるかどうかを判断するのに不要な情報

次のものを必要かどうか判断するのに不要な情報として概要には含めない。

- 画像 (<IMG...>)
- 音声などのハイパーテキストデータでないもの (HTTP の HEAD に対して content-type: として www/, html/text 以外を返すもの)
- 連絡先 (<ADDRESS...> .. </ADDRESS>)、
- HTML のヘッダ部 (<HEAD> ... </HEAD> または相当する部分

画像、音声は文字でないので概要にはなじまない。連絡先は内容に対するコメントの送り先を書くもので、内容には関係ない。 ヘッダ部は内容そのものではなく、ビュウアーへの情報や内容のタイトルなどである。

4.3 表現の密度

表現の密度を上げるために概要の HTML に次の変更を加える。

- <H1>ヘッダを<H2>ヘッダに変更する。
- リスト形式 (<OL>, <UL>, <DL>) をコンパクトリスト形式 (<OL COMPACT>, <UL COMPACT>, <DL COMPACT>) とする。

ヘッダはハイパーテキストの内容に構造を与える役割があるのでクライアントで表示されるときに文字の大きさを大きくし、下線を付ける。<H1>ヘッダは大見出しである。表現時にはヘッダの中では最大の文字サイズで表示される。 このため表現の密度を下げる。ハイパーテキストの内容の構造を残しながら、密度を上げるために、大見出しのみその下の見出しに変更した。 リスト形式はオプションでコンパクト形式を持つので、概要ではコンパクト形式とした。

5 評価

ここでは現状での概要作成サーバ、 概要作成サーバの出力についていくつかの点からの評価を行う。

表 1: http://www.wide.ad.jp/からリンクを辿る数を 10 としたときの転送量など

項目	回数	バイト数 (bytes)
Skipped Image	40	24837
Skipped Audio	0	0
Retrieved text	9	12683
Skipped Unknown	11	
Number of omitted click	10	
Number of parse error	0	
Number of HTTP error	0	

5.1 HTML のタグのカバレッジ

概要を作るためにハイパーテキストの記述言語である HTML の解析を行う。HTML のタグのいくつかを解析部が認識しないので、解析エラーとなる。ハイパーテキストの先頭で解析エラーとなると概要は全く作られない。途中のハイパーテキストの場合はそのリンクは辿られなかったのと同じとなる。解析部が認識しないタグがある理由は、HTML の有効な (De Fact)Standard がないためである。ネットワーク上の多くのハイパーテキストは HTML Spec 2.0[2] から見ると HTML Spec には存在しないタグ (例: <OWNER-NAME ...>) を使い、言語構造としては許されない入れ子構造 (例: <H1>は<BLOCKQUOTE>、<BODY> の中でしか使えないのだが, <UL>の中で使うなど) を使っているものが多い。 解析部を作る際に準拠すべき規格がない。これは WWW システムが発展途上のためともみなせるがこのましいことではない。

5.2 転送量, 操作数

概要を作成するために転送したハイパーテキストの転送量 (コネクションの数)、「不要と判断して」転送しなかった量 (開設しなかったコネクションの数)、出力量は概要を作成するハイパーテキストに依存する。表 1に http://www.wide.ad.jp/ からリンクを辿る数を 10 としたときの転送量などの情報を示す。この例に関しては次のことが言える。クライアントで画像を常に転送するような設定の場合と概要作成サーバを使った場合での転送量、TCP コネクションの開設数の比は 1/3($\approx$ 12683/(12683 + 24837)), 1/6($\approx$ 9/(40 + 9 + 11)) となる。TCP コネクションの開設数はコネクション開設時間として待ち時間に反映する。

5.3 出力は概要として使えるか?

概要というには量が多くなりがちである。その原因を述べる。

元のハイパーテキストのデータによるもの ハイパーテキストによってはいくつかの文字コード (JIS, EUC, S-JIS) のデータをリンクしていたり、日本語、英語のデータをリンクしているための場合もある。

概要作成サーバによるもの 物理制御情報、および、内容も残し過ぎているためとリンクを辿る範囲にたいして出力量からの負帰還がかかってないためである。残す情報を少なくした試みとして、必要な情報として各ハイパーテキストのタイトル (<TITLE> .. </TITLE>) のみを残し、ハイパーテキストでリンクを辿ると表示のときにネストするようなものを作成したがこれでは情報が落ち過ぎていた。[1]

5.4 リンクを辿る範囲は関連があるか?

リンクの種類と数だけで制限しているので関連があるとはいいがたい。

6 今後

今後は評価で不十分であることがわかった点について改良していく。ここでは改良の方法について述べる。

6.1 リンクを辿る範囲

数ではなく深さに着目してみる 先頭の HTML で同じサーバ上のリンクを辿る深さ、他のサーバへのリンクを辿る深さ、他のサーバに移ってからそのサーバ上のリンクを辿る深さ、他のサーバに移ってからまた別のサーバへのリンクを辿る深さ、それに同じサーバのリンクを優先するか、他のサーバはのリンクを優先するか、どちらも同じに扱うか を指定できるようにする。 この概要作成専用サーバでハイパーテキストによらない概要を作るのに最適な値があるかどうかを調べる。

内容に着目してみる タイトルの文字列に重なりがある場合のみリンクを辿るという概要作成専用サーバを試作してみる。タイトルの文字列が指定の文字列だとそのハイパーテキストは概要に含まないサーバも試作してみる。 この文字列としては What's New, History などを試行する。

6.2 判断に不要な情報

物理制御タグ (<FONT>, <CODE>) なども落としてみる。ヘッダー情報、それとリスト構造、それにパラグラフの最初の文などを残す概要作成専用サーバを試作してみる。

6.3 表示量/転送量

常に転送した量 (回数)、転送しなかった量 (回数)、それと表示量とのデータを取る。 使い安い表示を得る最小転送量となるかを検査する。

7 謝辞

本研究は日本科学技術情報センターからの委託により実施したものである。日本科学技術情報センターには、本研究の機会を与えて頂き感謝する。

8 付録

8.1 HTML のシンタックスあやまり

物理制御タグと論理制御タグとを入れ子にせづに使っている場合がおおい。タグのなかの KEYWORD =の後ろは必ずダブルクォート" で囲まれていけないのだがこれを省略している。開きタグと閉じタグとが組みになっているタグが開きタグだけである場合が多い。<HEAD>、<BODY> などが閉じていないことが多い。

8.2 日本語文字コード

HTML では<,>を特別扱いしている。 表示の際に<,>を表示するには&lt, &gt の記号を使う。多バイトコード中に<,>のコードを含む文字コードを HTML の中に混在させるのは HTML の設計思想に反しているので避けたほうがいい。

参考文献

[1] Tim Berners-Lee. The world wide web initiative. In *Proc. of Inet '93*, pp. DBC1-DBC5. Internet Society, 1993. http://info.cern.ch/pub/www/doc/INET93.ps.Z.

[2] T. Berners-Lee and D. Connolly. Hypertext markup language specification - 2.0.

[1] HTTP には内容に関してあるタグの情報だけを取ってくる機能がないので一度ハイパーテキスト全体をのデータをもってきてタイトルをとりだす。 したがって表示量/転送量 は悪い。