大規模ユーザ向け情報クリッピングシステム*

7L－8

岡本卓哉　村田英子　菅谷奈津子†
(株)日立製作所 情報・通信開発本部‡

1．はじめに

インターネットの普及に伴い、新聞社などによるオンラインニュースサービスが広がり始めた。これらのサービスで入手できる情報は膨大であり、全ての情報にユーザが目を通すことは、もはや不可能になっている。

このような状況において必須となるのは、入手した情報のうち、重要な情報だけをあらかじめ選別し通知する「クリッピング機能」を持ったシステムである。

筆者らは、1,000人以上の大規模なユーザから与えられた検索条件に対して、各条件に合致する情報をリアルタイムに検索し、該当するユーザに配送する情報クリッピングシステムを試作し、本機能の有効性について評価した。

2．システム概要

試作した情報クリッピングシステムの構成を図１に示す。

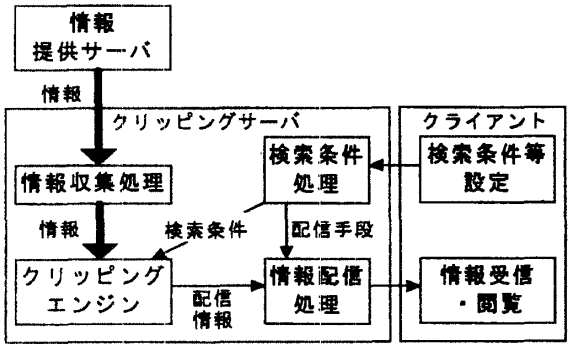

図１ 情報クリッピングシステムの構成

本システムでは、あらかじめクライアント（ユーザ）からの検索条件を受信し、クリッピングエンジンに登録する。クリッピングサーバに新しい情報が到着すると、その情報に対して、各ユーザが登録した検索条件との比較を行い、検索条件が成立するか否かを判定する。そして、検索条件が成立するユーザに対して、情報を配信する。

また、本システムでは、検索時にヒット位置を保存することにより、配信された情報を閲覧する際に、ヒット位置のハイライト表示を可能としている。

3．クリッピング処理方式

クリッピング処理は、図２に示したように、全ユーザの検索条件を基にしたオートマトンによる検索エンジンを利用して実現する。本オートマトンがターム取得状態に達した際に、各ユーザのタームカウンタをインクリメントする。そして、タームカウンタに対する論理条件処理により、各ユーザの検索条件を判定する。

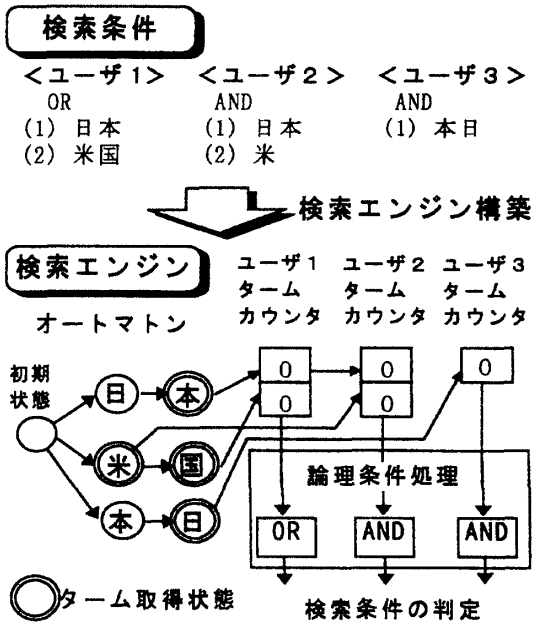

図２ 検索エンジンの構成

---

* Information Clipping System for a Large Number of Users.
† Takuya Okamoto, Eiko Murata, Natsuko Sugaya
‡ Information Systems R&D Division, Hitachi, Ltd.

上記のクリッピングエンジンに、検索対象の情報の先頭から順にスキャンして得られた文字を入力することで検索処理を実現する。本オートマトンがターム取得状態に達すると、本タームが検索条件中に含まれるユーザのタームカウンタに張られたリンクを辿り、全てのカウンタを1増加する。

論理条件に AND が設定されている場合、タームカウンタをチェックし、全てのカウンタの値が1以上であれば検索条件が成立する。また、OR が設定されている場合、少なくとも1つのカウンタが1以上であれば検索条件が成立する。

上記の処理により、1度テキストスキャンするだけで、複数のユーザが設定した検索条件を全てチェックすることができる。

4．評価実験

評価実験は、日立 WS 3050RX/335 (PARISC: 105MIPS)で実施した。1,000 人のユーザに対して、各々1つの検索タームからなる検索条件を設定した。そして、検索対象のテキストを変更することで、(1)オートマトンの処理、(2)ヒット文書番号格納、(3)ヒット位置格納の時間を求めた。

処理速度の評価は、以下の3通りのテキスト(長さは3,200ターム(=26,000文字))で行なった。

(I) 3,200 タームのいずれもヒットしない
(II)3,200 ターム中1タームが1,000人にヒット
(III)3,200 タームが 1,000 人にヒット

表1に上記テキストに対する処理時間の計測結果を示す。

表1 処理時間計測結果

評価テキスト	(I)	(II)	(III)
処理時間(ms)	40	4,000	23,000

(I)が 26,000 文字のテキストから検索タームを抽出する、オートマトンの処理時間、(II)から(I)を引いた時間が1,000人分のヒット文書番号の格納時間、(III)から(II)を引いた時間が、3,200,000箇所のヒット位置の格納時間となる。

表2にクリッピングエンジンの処理速度をまとめる。

表2 クリッピングエンジンの処理速度

(1)オートマトンの処理	650,000 バイト/秒
(2)ヒット文書番号格納	200 人/秒
(3)ヒット位置格納	160,000 ターム/秒

実際にユーザが検索条件式を設定した場合のモデルとして、入力されるテキストの平均長さを 2,000 文字、平均でヒットするユーザ数を 50 人 (ユーザ数の 5%) 、ヒットする検索タームの数を1人あたり 10 タームと仮定した。上記の仮定で得られた、1テキスト当たりのクリッピング処理時間は**表3**に示した通りである。

表3 クリッピング処理時間

処理内容	処理時間計算	処理時間
(1)オートマトン	2,000/650,000	0.003 秒
(2)文書番号格納	50/250	0.200 秒
(3)ヒット位置格納	50*10/160,000	0.003 秒
合　　計	－	0.206 秒

本クリッピングエンジンは、1,000ユーザの利用で、1秒間に5件程度の情報のクリッピング処理能力を持つことになる。**表3**から明らかなように、本システムで最も処理時間に影響を受けるのは、ヒットするユーザ数である。

5．まとめ

複数ユーザからの検索条件を一括処理する検索エンジンを開発し、情報クリッピングシステムに適用した。評価実験により 1,000 ユーザの検索条件を設定した場合で、1秒あたりに平均5件の情報を処理できることが確認できた。

6．参考文献

[1]畠山他：「ソフトウェアによるテキストサーチマシンの実現」, 情報処理学会情報学基礎研究報告, Vol. 92, No. 32, 25-2, pp. 19-25 (1992. 5)