D-018

大規模電子メールアーカイブシステム向け高速蓄積・検索方式

竹内 丈志　　加藤　守　　山岸 義徳　　中村　隆顕　　郡 光則
三菱電機株式会社 情報技術総合研究所

1. はじめに

近年、電子メールによる情報漏洩防止や内部統制を目的としたメールアーカイブ技術に注目が集まっている。しかし、大規模なメールアーカイブのサイズは 100TB まで達することがあり、メールの蓄積速度に追随できない、メールの検索を対話的に実行できない、という課題があった。我々は、100GB のメールを 1 日で蓄積し、1TB の大規模メールアーカイブを 1 秒で検索する、大規模電子メールアーカイブシステムを開発した。本稿では、その高速蓄積・検索方式と評価について報告する。

2. 電子メールアーカイブ

開発した電子メールアーカイブのシステム構成を図 1に示す。全文検索機能は索引用DBを管理し、索引の蓄積とメールに一意に割り当てられるID(メールID)の検索を行う。メール管理機能はメール用DBを管理し、メールの蓄積と取得を行う。

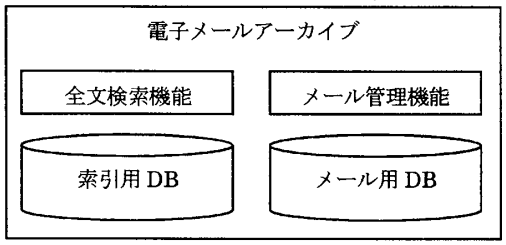

図 1　システム構成

蓄積処理では、添付ファイルを含むテキスト抽出を各メールに対して行い、索引用 DB に索引と属性情報およびメール ID を、メール用 DB にメールとメール ID を、それぞれ保存する。検索処理では、全文検索機能により任意のキーワードでメール ID を検索し、メール用 DB からメールを取得する。

3. 高速蓄積･検索方式

3.1. 蓄積・検索における課題

大企業で送受信されるメールは、数年間で 100TB にまで達する例があり、従来のメールアーカイブにはメールの蓄積または検索に以下のような課題があった。

- メールの蓄積速度に追随できない。
- メールの検索を対話的に実行できない。

メールアーカイブの検索の高速化には、索引が有効である。一方で、索引の蓄積には時間がかかるという課題がある。そこで、SISA[1][2][3]のディスクI/O技術および並列処理技術を用いた、高速蓄積・検索方式を適用した。

High Performance Load and Search Method of Mail Archive System.
Takeshi Takeuchi, Mamoru Kato, Takaaki Nakamura, Yoshinori Yamagishi, Mitsunori Kori
Information Technology R&D Center, Mitsubishi Electric Corporation.

3.2. 高速蓄積方式

蓄積処理では、メールアーカイブの蓄積サイズの増加に伴い、索引の更新処理に時間がかかることが課題であった。

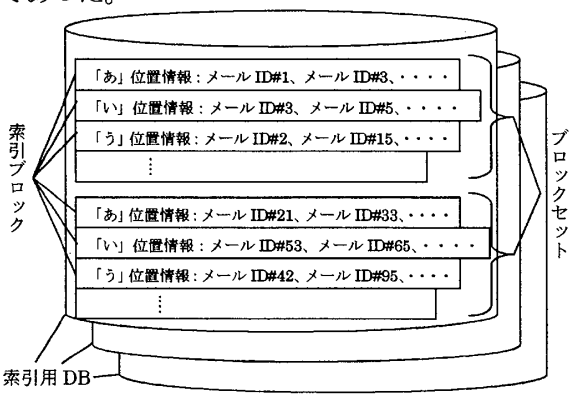

図 2　ブロック化 N-gram 索引形式

そこで、図 2のようにブロック化N-gram索引形式[4]で索引をブロック化し、全てのN-gramを含むように索引ブロックをまとめたブロックセット単位で管理する。また、各ブロックセットは複数のディスクへ分散配置する。索引の蓄積時には、任意のバッファサイズ単位で索引を一時ブロックセットとして追加していき、これらを検索バッファサイズ未満にマージして再度分配配置する最適化[5]を行う。このようにすることで、索引の更新処理を高速化し、高速蓄積を実現している。

3.3. 高速検索方式

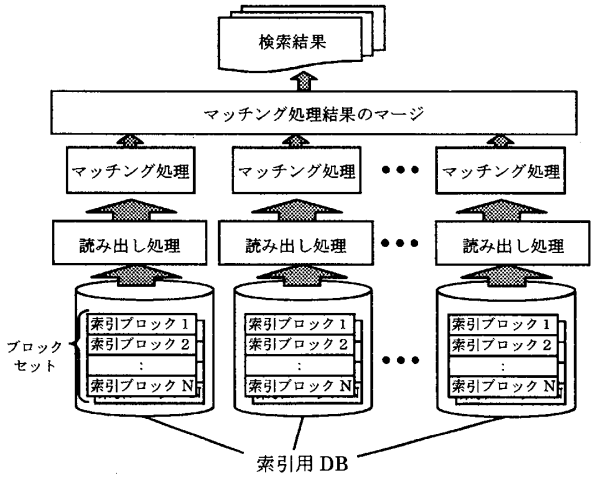

図 3　メール検索時のデータフロー

図 3にメール検索時のデータフローを示す。以下の方式により、高速検索を実現している[2] [4][5]。

- ブロック化N-gram索引形式[4]により、キーワ

ード文字列に対応した索引ブロックを、それぞれ一方向のシークでI/O転送。

- ディスクからの索引ブロックの読み出し、N-gram 索引のマッチング処理、マッチング処理結果のマージを、独立したスレッド単位で並列に実行。

4. 評価

本メールアーカイブの高速蓄積・検索方式の有効性を確認するため、それぞれの速度性能について評価した。

4.1. 評価システム構成

評価システムの構成を表 1に示す。

表 1 評価システム構成

OS	Windows 2003 Server x64 Enterprise Edition
CPU	Xeon MP 3.66GHz x 4
Memory	15.9 GB
HDD	台数:14 台, 回転速度:15000rpm, キャッシュ:16MB, 接続形態:Ultra SCSI 320

4.2. 蓄積性能評価

蓄積性能の評価には表 2の蓄積用評価データを用いた。

表 2 蓄積用評価データ構成

メール件数	1,093,485[件]
メールサイズ	99.65[GB]
テキストサイズ	22.40[GB]

表 2の蓄積用評価データを16.7時間で蓄積完了することができ、100GBのメールを 1 日で蓄積可能であることを確認した。また、図 4は蓄積時間とメールアーカイブの蓄積サイズの関係を示したものである。図 4から、メールアーカイブの蓄積サイズが増加しても、蓄積性能が低下しないことを確認できる。

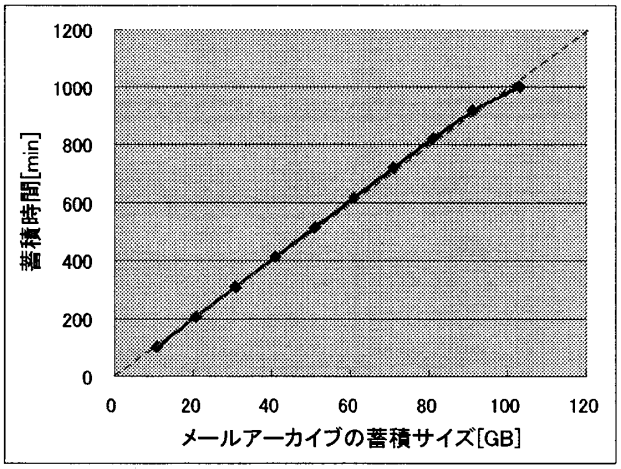

図 4 メールアーカイブの蓄積サイズと蓄積時間

4.3. 検索性能評価

検索性能の評価には表 3の検索用評価データを用いた。メール検索条件は 80 種類用意し、それぞれ 1 個の検索キーワードとした。文字数は 2～9、平均文字数は 3.7 である。

表 3 検索用評価データ

メール件数	10,934,850[件]
メールサイズ	996.5[GB]
テキストサイズ	224.0[GB]

平均検索時間の結果は 0.78[sec]となり、1TBの大規模メールアーカイブに対して 1 秒で検索可能であることを確認した。図 5は 80 種類のメール検索条件のうち、代表的な 9 種類のメール検索条件に対する結果を抜粋したものである。検索時間が 1 秒を超えるものは、出現頻度が高い文字を検索キーワードに含んでいるためと考えられる。

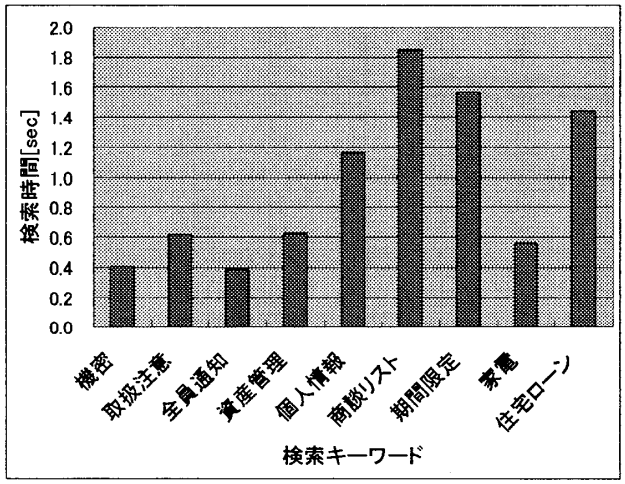

図 5 検索キーワードごとの検索時間

5. むすび

1 日 100GB のメールを蓄積でき、1TB の大規模メールアーカイブに対して 1 秒で高速検索が可能な、大規模電子メールアーカイブシステムを実現し、100TB の大規模メールアーカイブを処理できる十分な性能を確認できた。

参考文献

[1] 郡 他, 検索機能を備えたストレージシステムによる大規模並列全文検索, 信学技報, CPSY-2002-47, Aug-2002

[2] 清水 他, スケーラブルインテリジェントストレージによる大規模並列全文検索の実現, 第 64 回情処全国大会, 4ZA-4, Mar-2002

[3] 金子 他, スケーラブルインテリジェントストレージによる大規模並列全文検索の評価, 第 64 回情処全国大会, 4ZA-5, Mar-2002

[4] 山岸 他, n-gram 索引による大規模・並列全文検索方式－(1)実装と評価, 信学会ソサエティ大会, D-4-3, pp-21, Sep-2001

[5] 清水 他, n-gram 索引による大規模・並列全文検索方式－(2)索引の最適化, 信学会ソサエティ大会, D-4-4, pp-21, Sep-2001