
6

性能検証

6. 性能検証

6. 性能検証	1
6-1. 性能検証の考え方	2
6-1-1. 性能検証の目的	2
6-1-2. 性能測定環境	3
6-1-3. 性能測定で使用するデータ	4
6-1-4. GETAにおける検索特性	4
6-1-5. 性能測定パターン	5
6-1-6. 性能測定方法	7
6-2. 概念検索	8
6-2-1. 測定条件	8
6-2-2. 測定結果	9
6-2-3. 測定結果のまとめ	17
6-3. データマイニング	18
6-3-1. 測定条件	18
6-3-2. 測定結果	19
6-3-3. 測定結果のまとめ	27
6-3-4. 処理時間に対する改善策	28
6-4. 基礎数値の算出	29
6-5. 本番環境を想定したサイジング	30
6-6. データ蓄積	38
6-6-1. 蓄積データ概要	38
6-6-2. 蓄積データ	39
6-6-3. 測定方法	41
6-6-4. 測定結果と考察	43
6-6-5. 蓄積のまとめ	61
6-7. 性能検証のまとめ	62

6-1. 性能検証の考え方

6-1-1. 性能検証の目的

本章では、オンライン処理ならびに蓄積処理の性能検証を実施する。オンライン処理と蓄積処理における性能検証の目的をまとめる。

(1) オンライン処理

概念検索及びデータマイニングの検証ツール(オンライン処理)について、外部ユーザ利用環境(一般的なPCのスペック等)においても実用的なスピードが確保されるか、モデル検証環境での測定及び測定結果を用いた机上検証を行う。また、机上検証の結果から、性能目標値を満たすために必要なサーバ台数を見積る。性能検証の結果、目標性能を満たすことが難しい場合は、性能改善案について提示する。

(2) 蓄積処理

データ蓄積ツールの処理時間を測定することで、蓄積性能単価を算出する。また、性能単価から特許庁保有データ全件に対するデータ蓄積処理の予測時間の算出を行う。

6-1-2. 性能測定環境

性能測定を実施するマシン(PC とサーバ)とディスクの性能についてまとめる。

(1) マシン性能

性能検証を実施する測定環境のマシン性能を表 6-1-2-1 に示す。

表 6-1-2-1. マシン性能

#	マシン	用途	CPU	メモリ	OS/ブラウザ
1	AP サーバ	データ蓄積ツール実行環境 GETA の集約サーバ	Intel(R) Xeon(R) 3GHz x 2 (4Core)	2GB × 2	Red Hat Enterprise Linux 4.5
2	DB サーバ 1	GETA の分散サーバ	Intel(R) Xeon(R) 3GHz x 2 (4Core)	4GB × 2	Red Hat Enterprise Linux 4.5
3	DB サーバ 2	GETA の分散サーバ	Intel(R) Xeon(R) 3GHz x 2 (4Core)	4GB × 2	Red Hat Enterprise Linux 4.5
4	PC	クライアント PC	Intel(R) Core 2 Duo 1.8Hz (2Core)	2GB	Windows XP Internet Explorer 6.0

(2) ディスク性能

性能検証を実施する測定環境のディスク性能を表 6-1-2-2 に示す。

表 6-1-2-2. ディスク性能

#	装置	コントローラ	ホスト インタフェース	ディスクドライブ インタフェース	ディスク
1	Hitachi WMS100	最大キャッシュ容量 2GB (デュアルコントローラ)	FibreChannel (400 Mbyte/s)	SATA 1.5 Gb/s	250GB × 22 (7,200 rpm)

6-1-3. 性能測定で使用するデータ

性能測定で使用するデータの件数について表 6-1-3-1 に示す。

表 6-1-3-1. 性能検証で使用するデータの件数

#	名称	年数	範囲	件数	WAM サイズ
1	公開公報	2年	1994年から1995年	687,881件	2.1GB
2		4年	1994年から1997年	1,359,857件	4.4GB
3		8年	1994年から2001年	2,774,380件	9.3GB
4		全件(参考)	2009年2月時点	6,418,316件	21.5GB

6-1-4. GETA における検索特性

本検証では検索エンジンとして、GETA を用いている(GETA の詳細は 2-5-3 章参照)。ここでは、GETA の検索特性について述べる。

(1) 登録データの増加に関する特性

GETA を使用した概念検索は、公報から作成される WAM と呼ばれるマトリクス(行列)のデータに対して検索を行う。この WAM のサイズが大きいくほど、類似度算出のための計算量が多くなり検索時間が長くなる。

(2) 分散サーバの増加に関する特性

GETA は集約サーバと分散サーバに役割が分担されている。分散サーバが複数存在する場合、集約サーバは複数の分散サーバで処理された検索結果を統合する。そのため、分散サーバの数が多いほど1つの分散サーバにかかる負荷は減少するが、逆に、集約サーバへかかる負荷が大きくなる可能性がある。

6-1-5. 性能測定パターン

GETA の検索特性より、文献数の増加ならびに分散サーバ数の増加に対して性能測定を行う。

(1) 文献数の増加に関する性能検証

分散数を 4 に固定した状態で、登録データ件数を公開公報 2 年分、4 年分、8 年分と増加させ、公報登録件数を変化させた場合の検索時間を測定する。このパターンで性能測定を実施し、検索時間を取得することで、検索対象文献と検索時間の関係を分析する。

表 6-1-5-1. 検索対象文献件数の検証パターン

#	パターン	公報年数	文献数	分散サーバ数	DB サーバ数(※1)
1	検証パターン 1	公開公報 2 年分	687,881 件	4 分散	2 台(2 分散)
2	検証パターン 2	公開公報 4 年分	1,359,857 件	4 分散	2 台(2 分散)
3	検証パターン 3	公開公報 8 年分	2,774,380 件	4 分散	2 台(2 分散)

(※1)カッコ内は DB サーバ 1 台あたりの分散サーバ数

(2) 分散サーバ数の増加に関する性能検証

文献数を公開公報 2 年分に固定した状態で、分散サーバ数を 1、2、4、8 と増加させ、分散サーバ数を変化させた場合の検索時間を測定する。このパターンで、分散サーバ数の変化がどの程度性能に影響するかを測定する。DB サーバと分散サーバの構成を図 6-1-5-1 に示す。

表 6-1-5-2. 分散サーバ数の検証パターン

#	パターン	公報年数	文献数	分散サーバ数	DB サーバ数(※1)
1	検証パターン 4	公開公報 2 年分	687,881 件	1 分散	1 台(1 分散)
2	検証パターン 5	公開公報 2 年分	687,881 件	2 分散	1 台(2 分散)
3	検証パターン 6	公開公報 2 年分	687,881 件	4 分散	1 台(4 分散)
4	検証パターン 7	公開公報 2 年分	687,881 件	8 分散	2 台(4 分散)

(※1)カッコ内は DB サーバ 1 台あたりの分散サーバ数

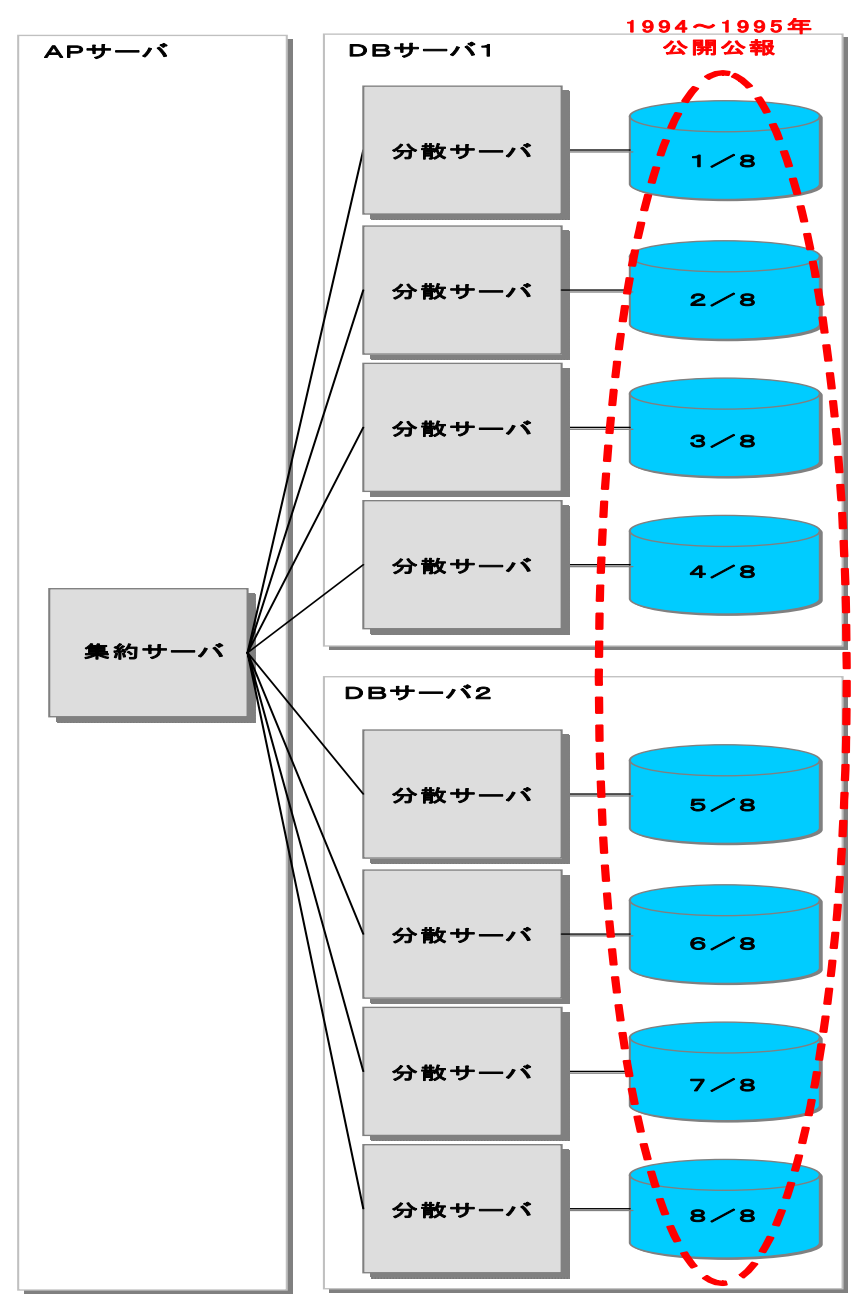
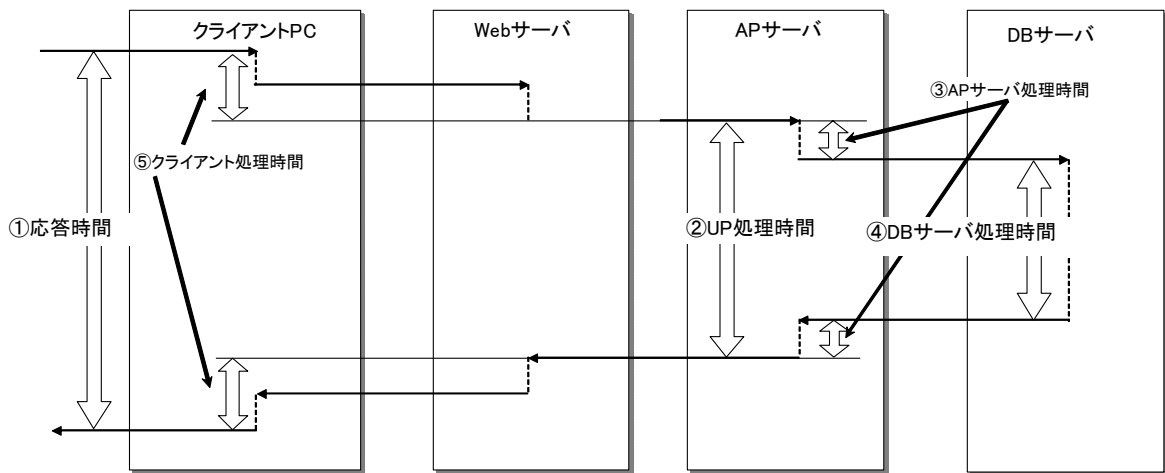


図6-1-5-1. DBサーバと分散サーバの関係

6-1-6. 性能測定方法

(1) 処理概要

概念検索ならびにデータマイニングの処理フローを図6-1-6-1に示す。



※ネットワーク上の処理時間は、PC・サーバ内の時間に含まれている。

図6-1-6-1. 処理フロー

(2) 処理時間の測定範囲

処理時間の測定範囲を表6-1-6-1に示す。

表6-1-6-1. 測定範囲

#	項目	測定範囲
1	①応答時間	検索を実行してから、結果が表示されるまでの時間 ヒットした文献一覧と特徴語を取得し表示する。
2	②UP 処理時間	UP 実行時間。
3	③DB サーバ処理時間	DB サーバの CPU 処理時間
4	④AP サーバ処理時間	AP サーバの CPU 処理時間
5	⑤クライアント処理時間	PC および Web サーバの処理時間。応答時間からサーバの UP 処理時間を引いたもの

(3) リソース使用状況の測定範囲

リソース使用状況を測定する項目を表6-1-6-2に示す。

表6-1-6-2. リソース使用状況の測定項目

#	測定項目
1	クライアントPC CPU利用率
2	クライアントPC DISK利用率
3	クライアントPC メモリ利用率
4	サーバ CPU利用率(CPU時間)
5	サーバ IO利用率
6	サーバ メモリ利用量

6-2. 概念検索

6-2-1. 測定条件

概念検索の検索条件を表6-2-1-1に示す。

表6-2-1-1. 概念検索の検索条件

#	画面入力条件	入力内容	備考
1	本願	特開平 06-111111 特開平 06-333336 特開平 06-333339 特開平 07-000001 特開平 07-330010	左記の5つの本願についてそれぞれ測定する。
2	検索方式	特定箇所クエリ指定	フリーオペレーションからのアンケートで最も多く選択された条件を使用する。
3	検索クエリ	請求項	
4	検索対象	全文	
5	制限条件	本願のテーマ	
6	特徴語数	70	
7	取得文献数	1000	今回の概念検索のチューニングで使用している値を使用する。 現行の文献一覧表示と同じ件数を設定する。

6-2-2. 測定結果

ここでは、概念検索の性能測定結果についてまとめる。測定結果より、どの本願についても同様の傾向が見られるため、ここでは、特開平 06-111111 の検索結果を元に結果を整理する。なお、検索時間は 3 回の検索の平均値を使用している。

(1) サーバ処理時間

(a) 文献数増加時の処理時間の推移

文献数増加時の処理時間の推移を図 6-2-2-1 に示す。なお、ここで示す CPU 時間は DB サーバの CPU 時間であり、DB サーバ 1 と DB サーバ 2 のすべてのコアの合計値である。測定結果より、文献数に比例して処理時間が長くなることが分かる。

処理時間の主な変動要因は、DB サーバにおける CPU 時間である。図 6-2-2-1 より文献数が 2 倍になると、DB サーバの CPU 時間も 2 倍となることが分かる。

また、AP サーバの CPU は 1 コアが利用され、文献数の増加に依存せず、CPU 時間はほぼ一定である。

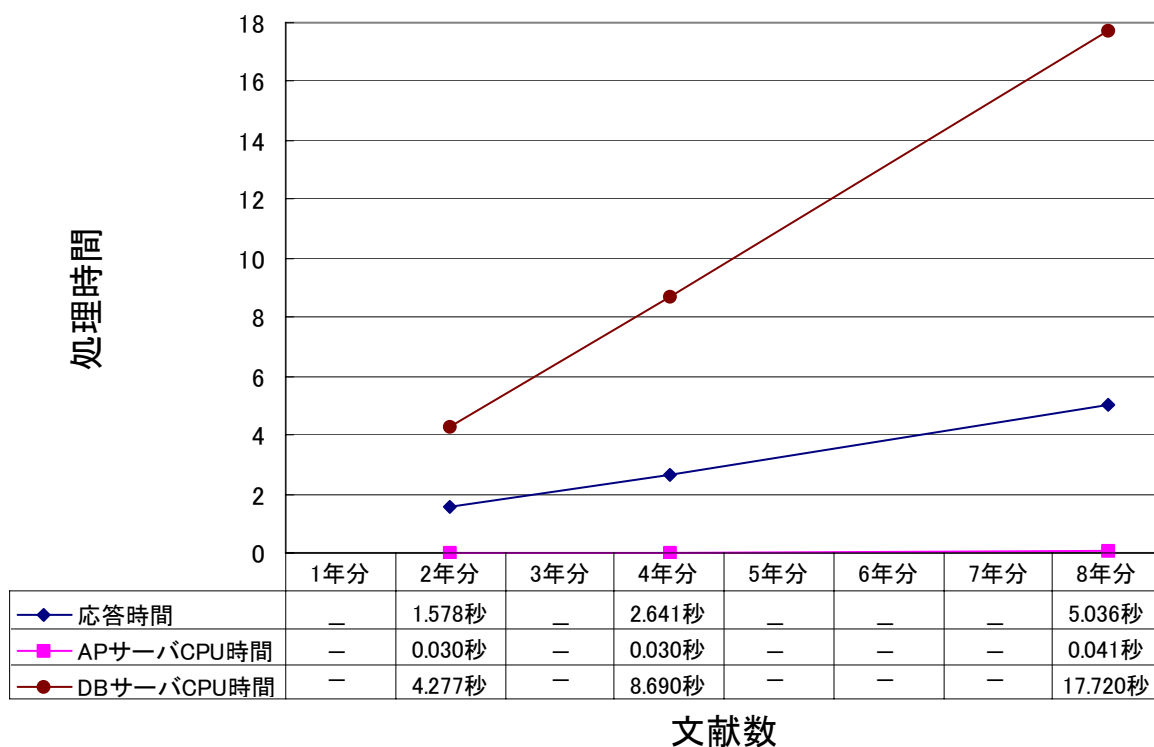


図 6-2-2-1. 文献数増加時の処理時間 (特開平 06-111111)

表 6-2-2-1. 文献数増加時のコアごとの CPU 処理時間 (特開平 06-111111)

#	文献数	AP サーバ				AP サ ーバ 合計	DB サーバ 1				DB サーバ 2				DB サ ーバ合 計
		コア 0	コア 1	コア 2	コア 3		コア 0	コア 1	コア 2	コア 3	コア 0	コア 1	コア 2	コア 3	
1	2年分	0.030	0.000	0.000	0.000	0.030	1.051	0.000	1.030	0.050	1.051	0.050	0.000	1.046	4.277
2	4年分	0.010	0.020	0.000	0.000	0.030	2.173	2.165	0.000	0.041	2.150	0.050	0.000	2.111	8.690
3	8年分	0.041	0.000	0.000	0.000	0.041	4.405	0.040	4.435	0.020	4.416	0.000	0.000	4.414	17.720

(b) 分散サーバ数増加時の処理時間の推移

分散サーバ数増加時の処理時間の推移を図6-2-2-2に示す。ここで示す CPU 時間は DB サーバの CPU 時間であり、DB サーバ 1 と DB サーバ 2 のすべてのコアの合計値である。測定結果より、分散サーバ数が変わっても、DB サーバの総 CPU 時間が一定であることが分かる。DB サーバの CPU 利用状況から、分散サーバ数が DB サーバのコア数以下の場合、分散サーバ数に応じて CPU のコアが使われることから、分散サーバ数が 2 倍になると、1 コアあたりの CPU 時間は 1/2 になることが分かる。

また、AP サーバの CPU は 1 コアが利用され、分散サーバ数の増加に依存せず、CPU 時間はほぼ一定である。

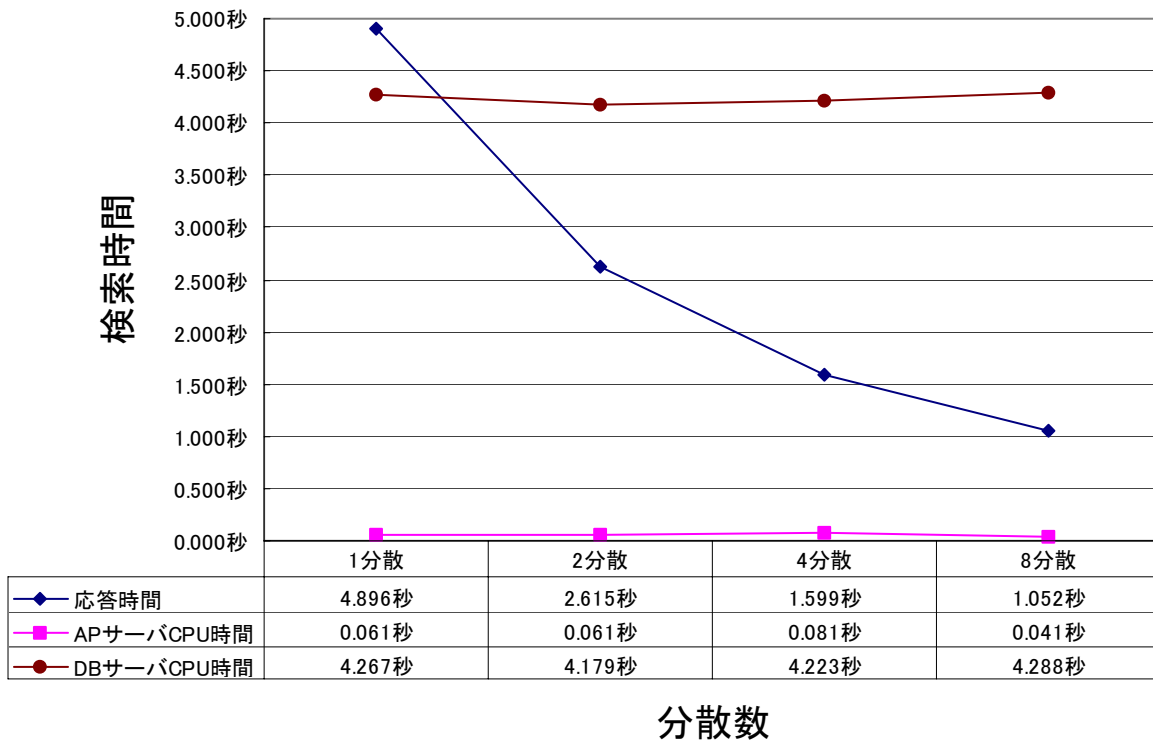


図6-2-2-2. 分散サーバ数増加時の処理時間 (特開平 06-111111)

表6-2-2-2. 分散サーバ数増加時のコアごとの CPU 処理時間 (特開平 06-111111)

#	分散数	AP サーバ				AP サ ーバ 合計	DB サーバ 1				DB サーバ 2				DB サ ーバ 合計
		コア 0	コア 1	コア 2	コア 3		コア 0	コア 1	コア 2	コア 3	コア 0	コア 1	コア 2	コア 3	
1	1分散	0.010	0.051	0.000	0.000	0.061	4.206	0.030	0.000	0.031	0.000	0.000	0.000	0.000	4.267
2	2分散	0.010	0.051	0.000	0.000	0.061	2.076	0.041	2.063	0.000	0.000	0.000	0.000	0.000	4.179
3	4分散	0.020	0.061	0.000	0.000	0.081	1.053	1.073	1.063	1.035	0.000	0.000	0.000	0.000	4.223
4	8分散	0.021	0.010	0.010	0.000	0.041	0.535	0.530	0.541	0.520	0.540	0.531	0.520	0.571	4.288

(2) クライアント処理時間

(a) 文献数増加時の処理時間の推移

文献数増加時のクライアント処理時間を図6-2-2-3に示す。クライアントPCの処理時間はPC側の応答時間とサーバ側のUP処理時間の差分である。文献数増加に依存せず、クライアントPCの処理時間はほぼ一定であることが分かる。

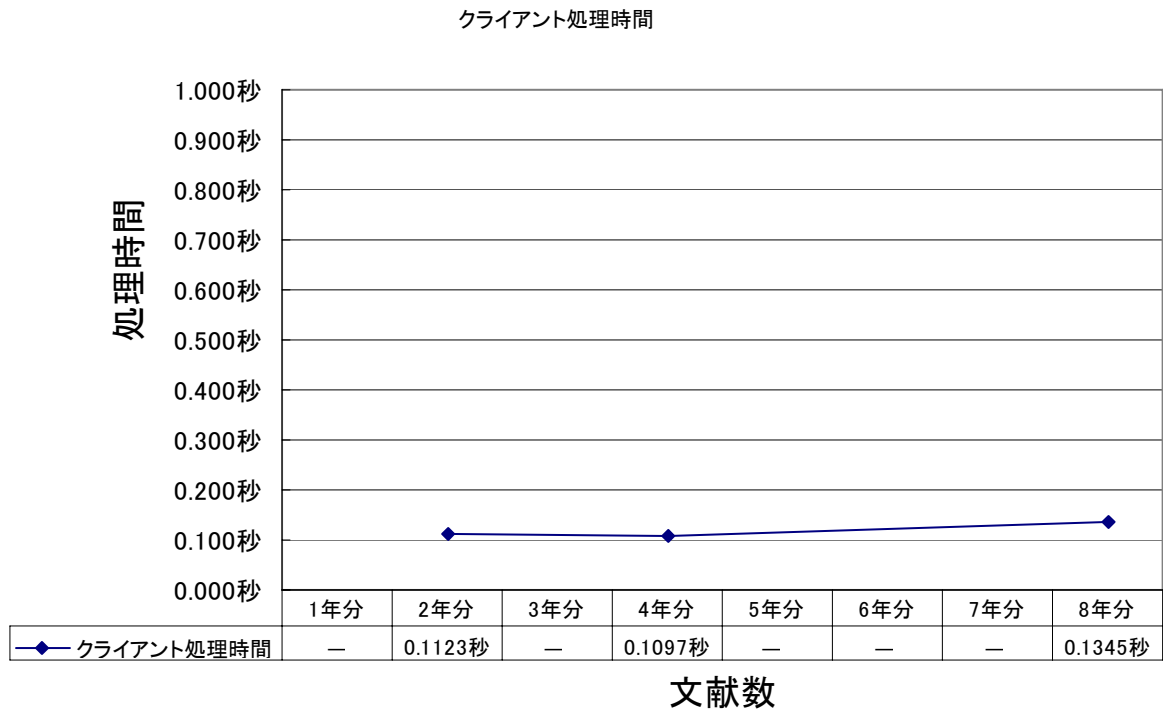


図6-2-2-3. 文献数増加時のクライアント処理時間 (特開平 06-111111)

(b) 分散サーバ数増加時の処理時間の推移

分散サーバ数増加時のクライアント処理時間を図6-2-2-4に示す。クライアント PC の処理時間は PC 側の応答時間とサーバ側の処理時間の差分である。分散サーバ数に依存せず、クライアント PC の処理時間はほぼ一定であることが分かる。

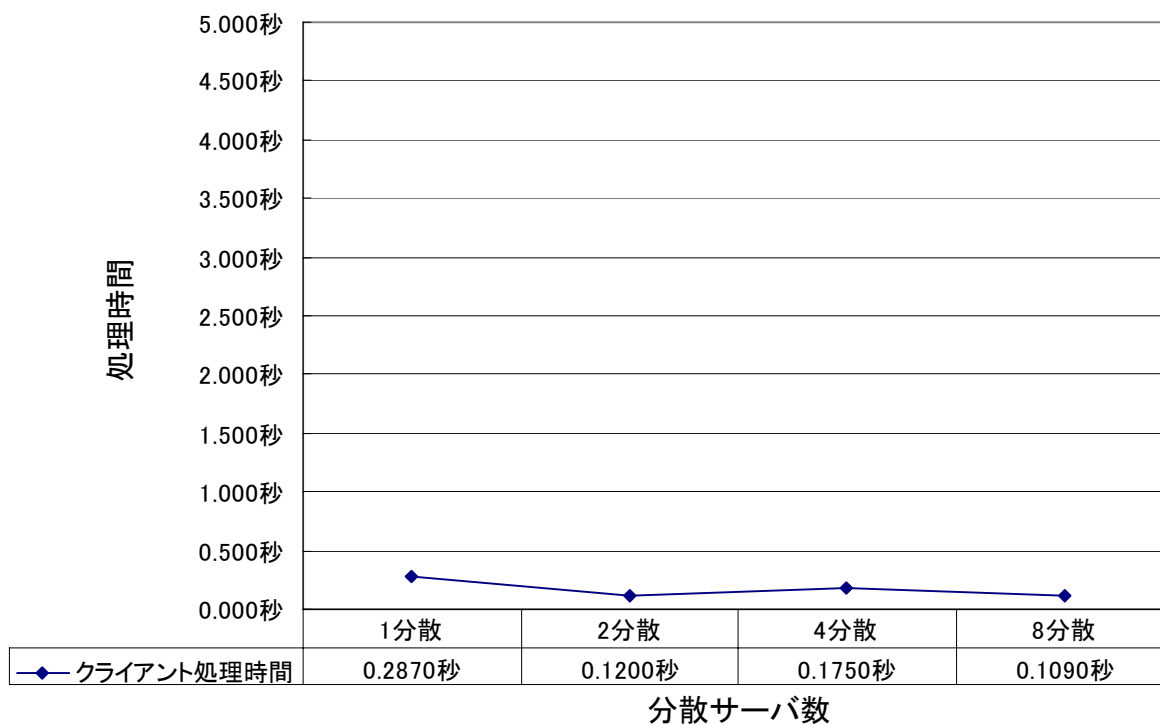


図6-2-2-4. 分散サーバ数増加時のクライアント処理時間 (特開平 06-111111)

(3) サーバリソース使用状況

(a) AP サーバ

文献数や分散サーバ数に依存せず、APサーバのリソース使用状況はほぼ同様であった。CPU利用率、メモリ利用量ともに低く、I/Owaitも発生しない(図6-2-2-5参照)。なお、CPUは4コアのうち1コアのみが使用される。

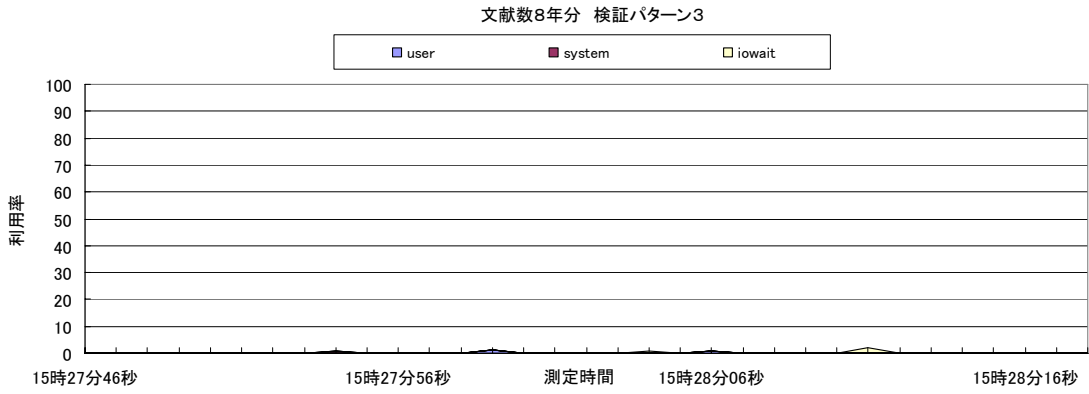


図6-2-2-5. 検証パターン3のAPサーバCPU利用率(コア1)

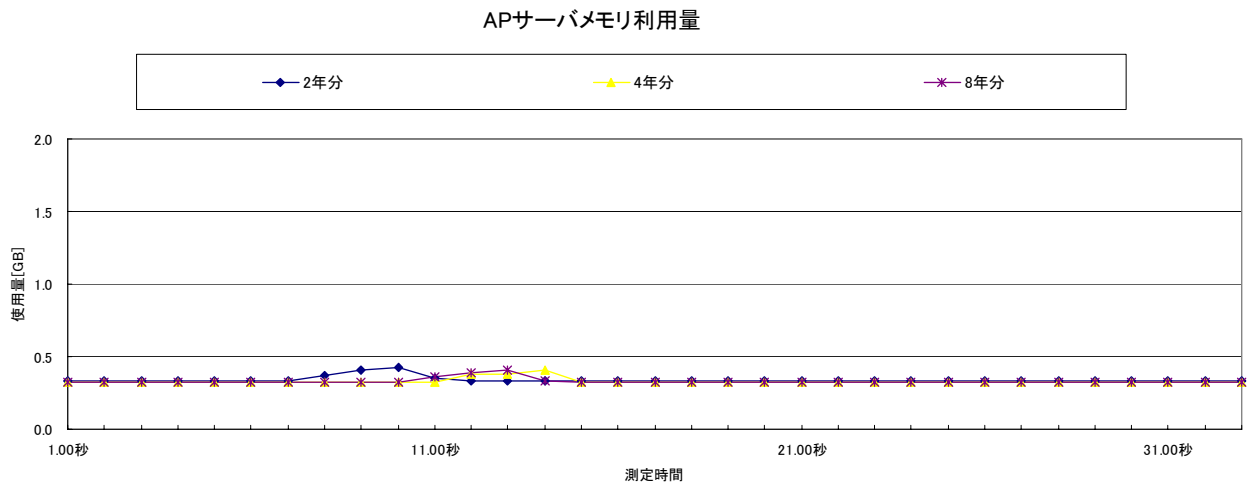


図6-2-2-6. 検証パターン3のAPサーバメモリ利用量

(b) DB サーバ(文献数増加時)

文献数増加時の DB サーバの CPU 利用状況を図 6-2-2-7 に示す。また、文献数増加時のメモリ利用状況を図 6-2-2-8 に示す。処理時間の推移でも述べたとおり、文献数増加に比例して、1 コアあたりの CPU 利用率は増加していることが分かる。なお、いずれの検証パターンにおいても、WAM はメモリ上に展開され、I/Owait は発生していない(図 6-2-2-9 参照)。メモリ利用量は、文献数にほぼ比例し、8 年分使用時で約 10%程度となる。

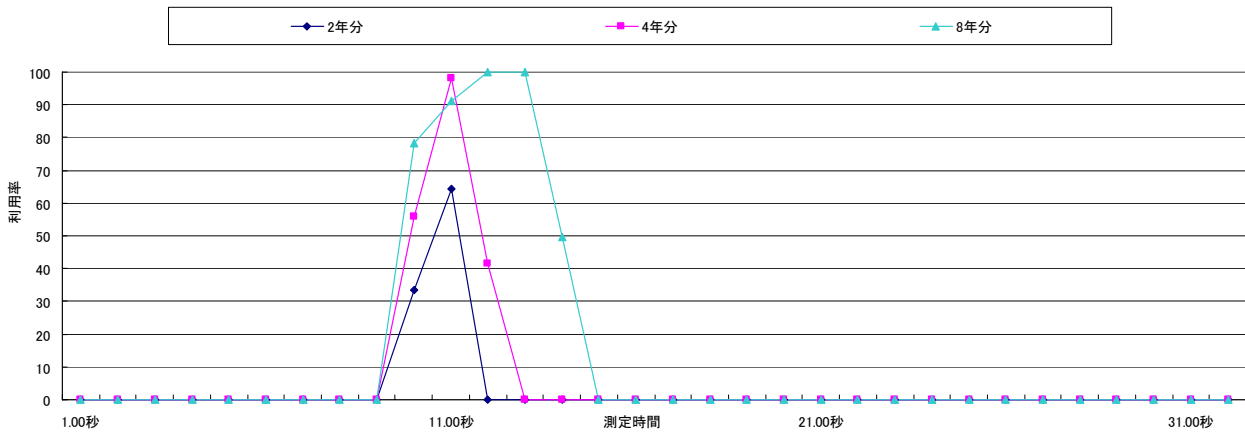


図 6-2-2-7. 文献数増加時の DB サーバの CPU 利用率 (特開平 06-111111)
(1 コアあたりの CPU 利用率)

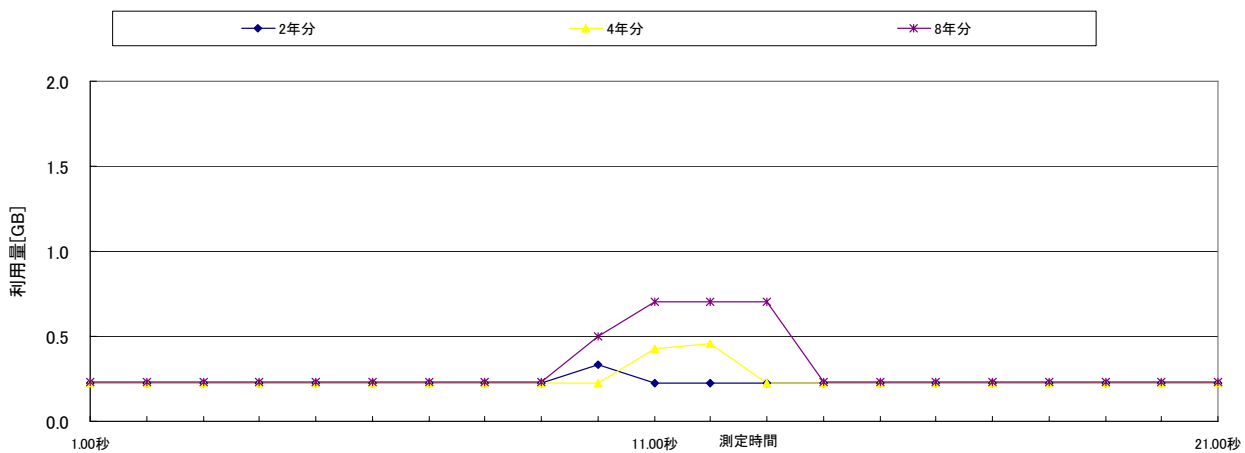


図 6-2-2-8. 文献数増加時の DB サーバのメモリ利用量 (特開平 06-111111)

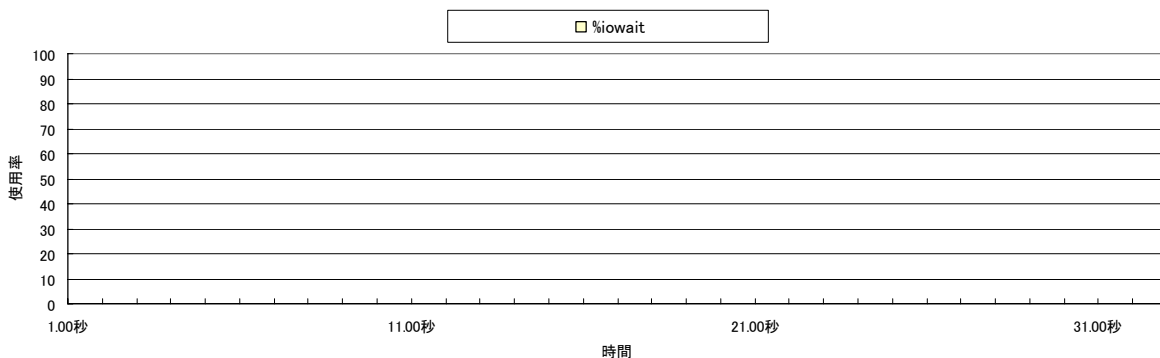


図 6-2-2-9. DB サーバの I/Owait 利用率 (8 年分)

(c) DB サーバ(分散サーバ数増加時)

分散サーバ数増加時の DB サーバの CPU 利用状況を図 6-2-2-10 に示す。また、分散サーバ数増加時のメモリ利用状況を図 6-2-2-11 に示す。分散サーバ数増加に比例して、1 コアあたりの CPU 利用率は減少していることが分かる。なお、いずれの検証パターンにおいても、WAM はメモリ上に展開され、I/Owait は発生していない(図 6-2-2-12 参照)。また、文献数が同じであれば、メモリ利用量の合計は分散数に関係なくほぼ一定となる。DB サーバが 1 台の場合、メモリ利用量は約 5%、DB サーバが 2 台の場合は 1 台の半分程度になることが分かる。

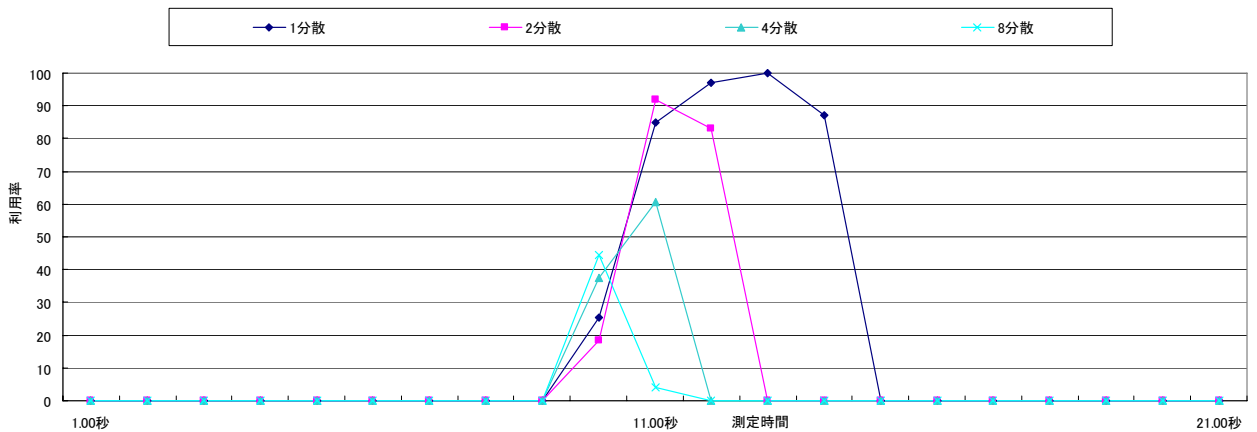


図 6-2-2-10. 分散数増加時の DB サーバの CPU 利用率 (特開平 06-111111)
(1 コアあたりの CPU 利用率)

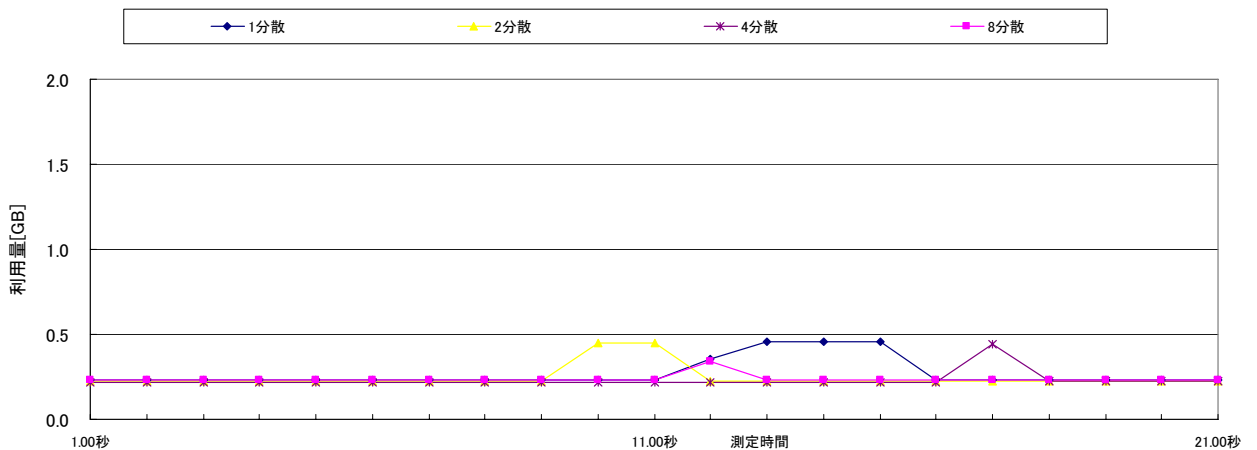


図 6-2-2-11. 分散サーバ数増加時の DB サーバ 1 のメモリ利用量 (特開平 06-111111)

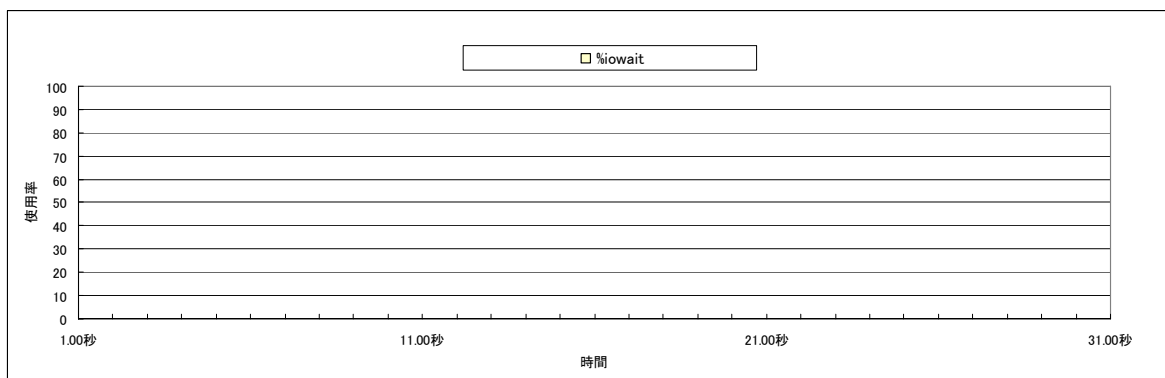


図 6-2-2-12. DB サーバの I/Owait 利用率 (8 分散)

(4) クライアントPCリソース使用状況

文献数や分散サーバ数に依存せず、クライアントPCのリソース使用状況はほぼ一定であった(図6-2-2-13)。性能検証で使用した一般的なクライアントPCレベルであれば、本調査で利用した概念検索は問題なく稼動できることが分かる。

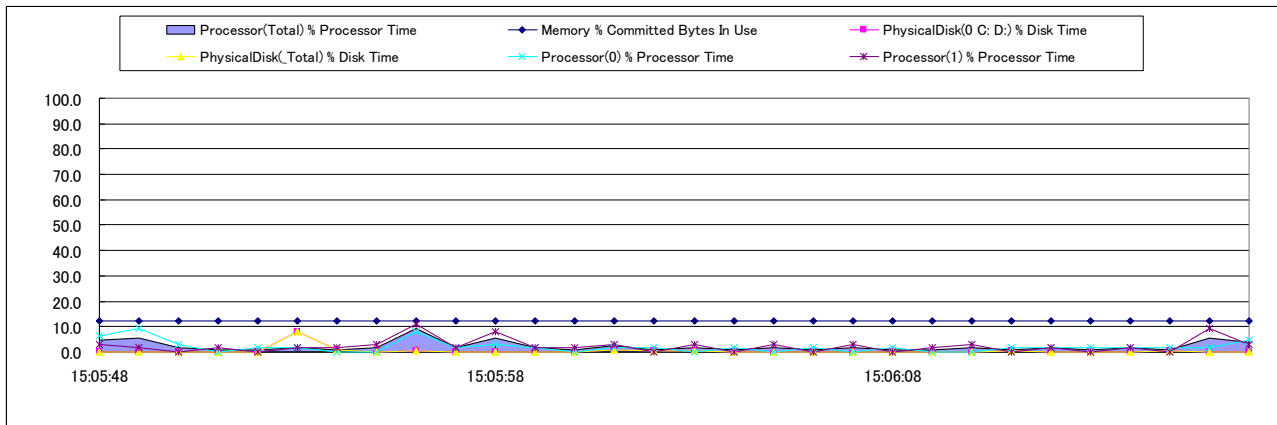


図6-2-2-13. クライアントPCリソース利用率 (特開平 06-111111)

6-2-3. 測定結果のまとめ

概念検索の測定結果を表6-2-3-1にまとめる。

表6-2-3-1. 概念検索の測定結果

#	マシン	測定結果	参照先
1	AP サーバ	文献数、分散サーバ数に依存せず、AP サーバの処理時間はほぼ一定である	図6-2-2-1 図6-2-2-2
2		CPUは1コアのみ負荷がかかるが、その利用率は低い	表6-2-2-1 表6-2-2-2
3		I/Owait やメモリなどのボトルネックは発生しない	図6-2-2-5 図6-2-2-6
4	DB サーバ	文献数に比例し、DB サーバの総 CPU 時間は増加する	図6-2-2-1
5		分散サーバ数に比例して、負荷がかかる CPU コア数は増加する	表6-2-2-2
6		分散サーバ数が増加すると、CPU1 コアあたりの CPU 時間は減少する	表6-2-2-2
7		メモリのボトルネックは発生しない	表6-2-2-8 表6-2-2-11
8		WAM はメモリ上に展開され、I/Owait のボトルネックは発生しない	表6-2-2-9 表6-2-2-12
9	クライアント PC	文献数、分散サーバ数に依存せず、クライアント PC の利用率はほぼ一定である	表6-2-2-13
10		CPU 利用率、メモリ利用量は小さく、一般的な PC でも十分に処理可能である	表6-2-2-13

6-3. データマイニング

6-3-1. 測定条件

データマイニングの測定条件を表6-3-1-1に示す。

表6-3-1-1. 検証用マイニング条件

#	画面入力条件	入力内容	条件選択理由
1	ワード項目	装置 設備 方法 製造 機器	11/11~12/12 日までのマイニング履歴データからマイニング条件に多く検索されたワードを選び、そのワードを関連語辞書から抽出した単語
2	観点	関連ワード (明細書から)	フリーオペレーションからのアンケートで最も多く選択された条件
3	文献数	200	
4	項目数	35	
5	出力特性	均等頻度	

6-3-2. 測定結果

ここでは、データマイニングの測定結果についてまとめる。なお、処理時間は3回の測定の平均値を使用している。

(1) サーバ処理時間

(a) 文献数増加時の処理時間の推移

文献数増加時の処理時間の推移を図6-3-2-1に示す。なお、ここではCPU時間はDBサーバ、APサーバそれぞれのコアの合計値である(DBサーバはDBサーバ1とDBサーバ2の合計値)。文献数の増加に比例して、APサーバとDBサーバのCPU時間が増加することが分かる。APサーバのCPU時間が増加する原因は文献数が増加することにより、スペクトル表示のために処理を行う特徴語数が増加するためである。(詳細は後述する)

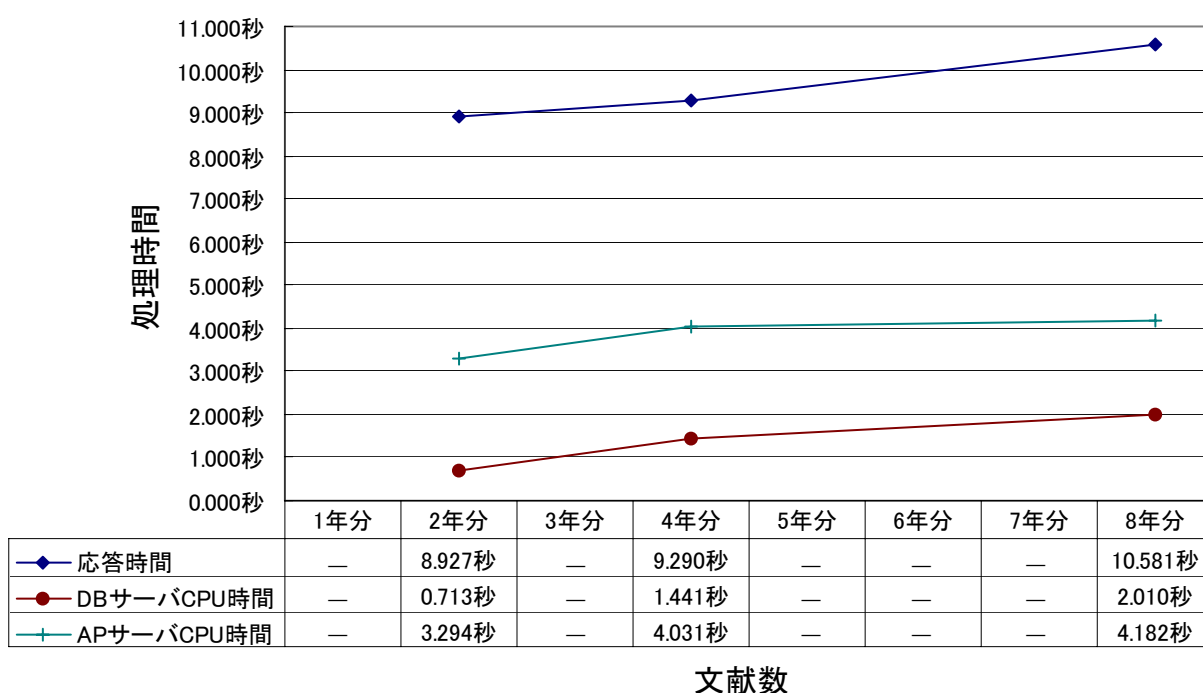


図6-3-2-1. 文献数増加時の処理時間

表6-3-2-1. 文献数増加時のコアごとのCPU処理時間

#	文献数	APサーバ				APサーバ合計	DBサーバ1				DBサーバ2				DBサーバ合計
		コア0	コア1	コア2	コア3		コア0	コア1	コア2	コア3	コア0	コア1	コア2	コア3	
1	2年	2.980	0.020	0.092	0.203	3.294	0.174	0.163	0.010	0.031	0.174	0.162	0.000	0.000	0.713
2	4年	2.976	0.738	0.194	0.122	4.031	0.377	0.384	0.000	0.010	0.316	0.030	0.020	0.303	1.441
3	8年	3.066	0.810	0.184	0.122	4.182	0.510	0.030	0.000	0.480	0.490	0.020	0.000	0.480	2.010

(b) 分散サーバ数増加時の処理時間の推移

分散サーバ数増加時の処理時間の推移を図6-3-2-2に示す。ここではCPU時間はDBサーバ、APサーバそれぞれのコアの合計値である(DBサーバはDBサーバ1とDBサーバ2の合計値)。測定結果より、分散数が変わっても、DBサーバの総CPU時間は一定であることが分かる。DBサーバのCPU利用状況から、分散サーバ数がDBサーバのコア数以下の場合、分散サーバ数に応じてCPUのコアが使われることから、分散サーバ数に応じて、1コアあたりのCPU時間は短くなることが分かる。

また、APサーバのCPUは1コアのみ利用され、分散サーバ数の増加に依存せず、3~4秒となる。概念検索と比較して、データマイニングは検索前後のUPの処理時間が重く、APサーバ上のUPの処理でCPUが使われる傾向にある。

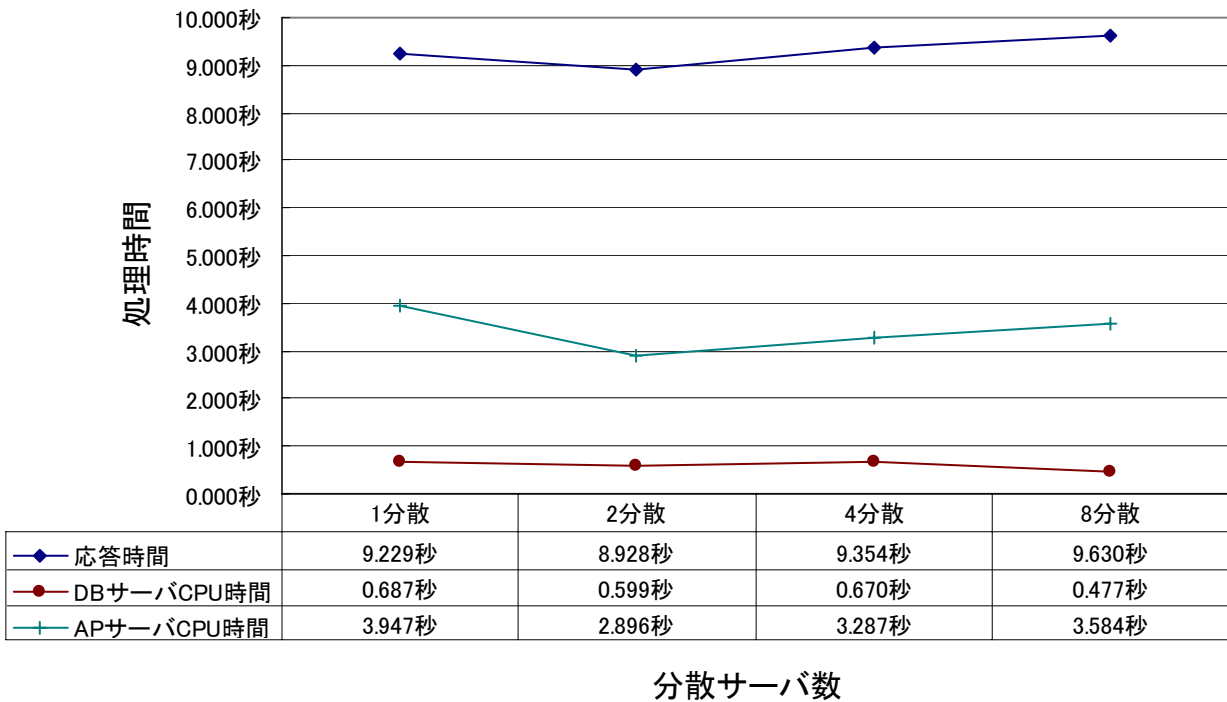


図6-3-2-2. 分散サーバ数増加時の処理時間

表6-3-2-2. 分散サーバ数増加時のコアごとのCPU処理時間

#	分散数	APサーバ				APサーバ合計	DBサーバ1				DBサーバ2				DBサーバ合計	
		コア0	コア1	コア2	コア3		コア0	コア1	コア2	コア3	コア0	コア1	コア2	コア3		
1	1分散	3.947	0.000	0.000	0.000	3.947	0.687	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.687
2	2分散	2.896	0.000	0.000	0.000	2.896	0.306	0.000	0.293	0.000	0.000	0.000	0.000	0.000	0.000	0.599
3	4分散	3.287	0.000	0.000	0.000	3.287	0.184	0.162	0.163	0.162	0.000	0.000	0.000	0.000	0.000	0.670
4	8分散	3.584	0.000	0.000	0.000	3.584	0.071	0.061	0.051	0.070	0.061	0.041	0.061	0.061	0.477	

クライアントPC処理時間

(a) 文献数増加時の処理時間の推移

文献数増加時のクライアントPC処理時間を図6-3-2-3に示す。クライアントPCの処理時間はPC側の応答時間とサーバ側の処理時間の差分である。文献数増加に比例して増加する。スペクトル表示処理で表示する特徴語が増えるためである。(詳細は後述する)

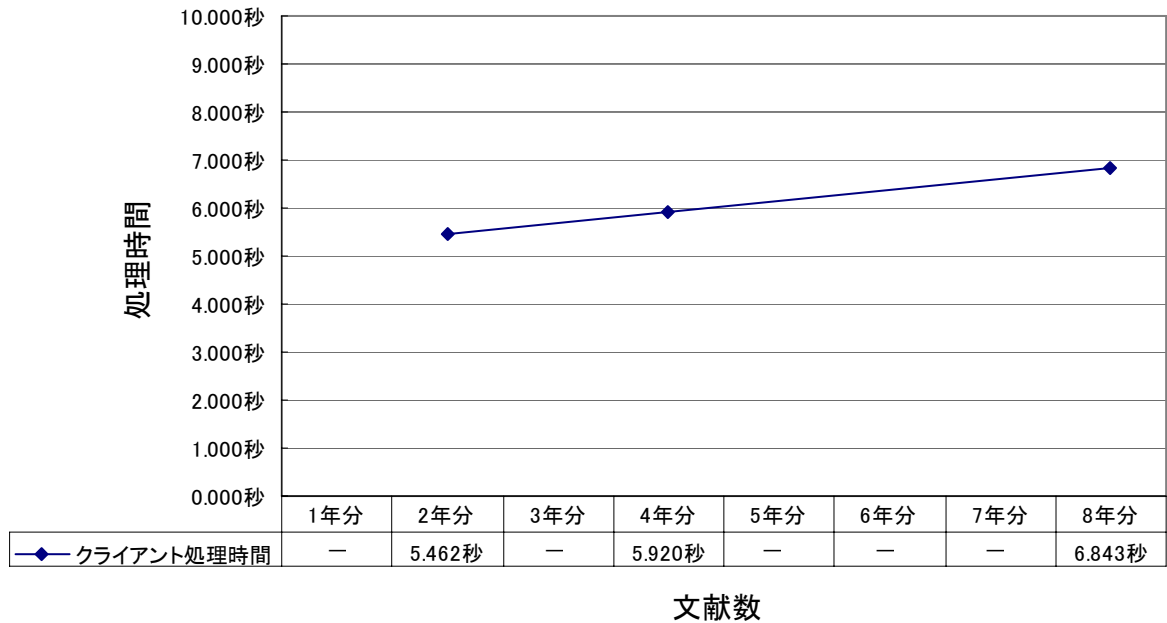


図6-3-2-3. 文献数増加時のクライアント処理時間

(b) 分散サーバ増加時の処理時間の推移

分散サーバ数増加時のクライアントPC処理時間を図6-3-2-4に示す。クライアントPCの処理時間はPC側の応答時間とサーバ側の処理時間の差分である。分散サーバ数に依存せず、クライアントPCの処理時間は一定であることが分かる。

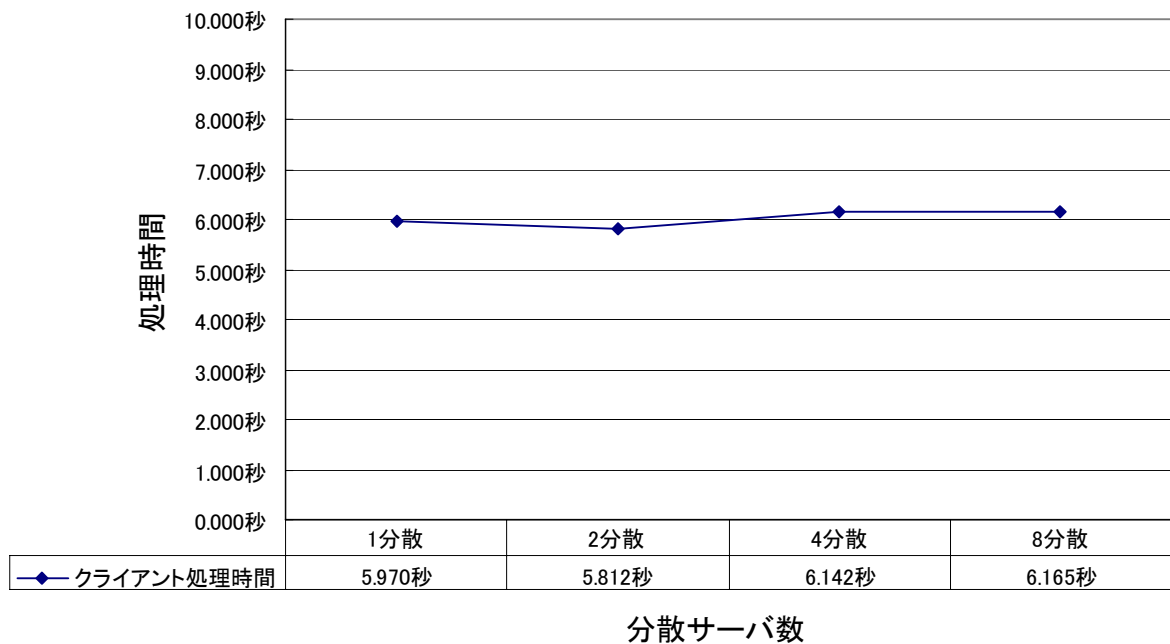


図6-3-2-4. 分散サーバ数増加時のクライアント処理時間

(3) サーバリソース使用状況

(a) AP サーバ(文献数増加時)

AP サーバの CPU 利用率を図 6-3-2-5 に示す。図は AP サーバの 4 コアのうち、実際に CPU に負荷がかっていた 1 コアのみでのグラフである。概念検索と比較して、AP サーバの CPU 利用率が高いことが分かる。これは、データマイニングでは GETA に対する検索よりも、AP サーバ上の UP の処理が重いからである。また、メモリ利用量はほぼ一定となる (図 6-3-2-6 参照)。なお、いずれの検証パターンにおいても、I/Owait は発生していない(図 6-3-2-7 参照)。

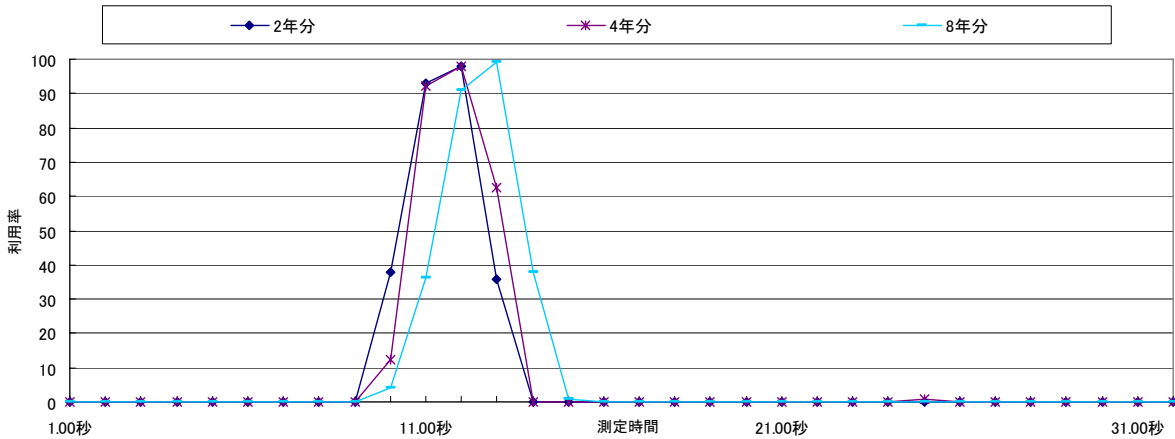


図 6-3-2-5. 文献数増加時の AP サーバの CPU 利用率
(1 コアあたりの CPU 利用率)

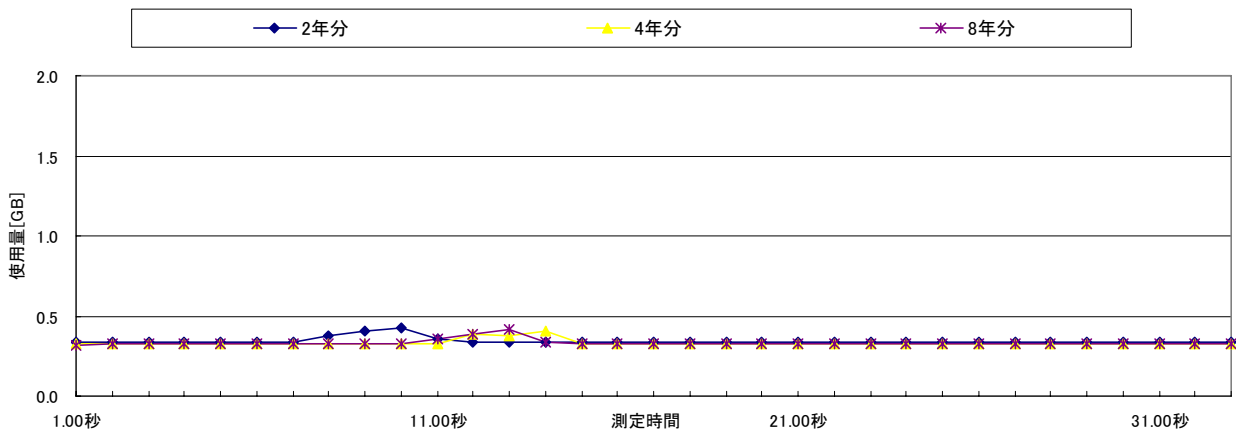


図 6-3-2-6. 文献数増加時の AP サーバのメモリ利用量

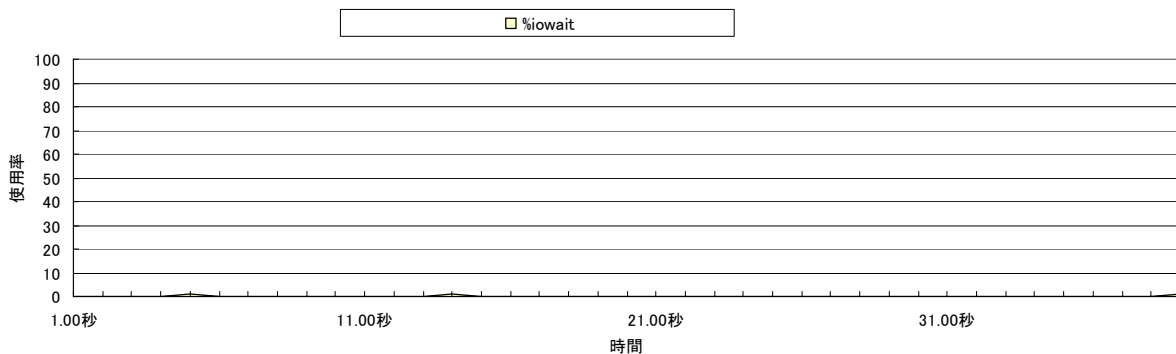


図 6-3-2-7. AP サーバの I/Owait 利用率 (8 年分)

(b) DB サーバ(文献数増加時)

文献数増加時の DB サーバの CPU 利用率を図 6-3-2-8 に示す。また、メモリ使用状況を図 6-3-2-9 に示す。文献数に比例して、CPU1 コアあたりの CPU 利用率が増加していることが分かる。メモリ利用量はほぼ一定となる。なお、いずれの検証パターンにおいても、WAM はメモリ上に展開され、I/Owait は発生していない(図 6-3-2-10 参照)。

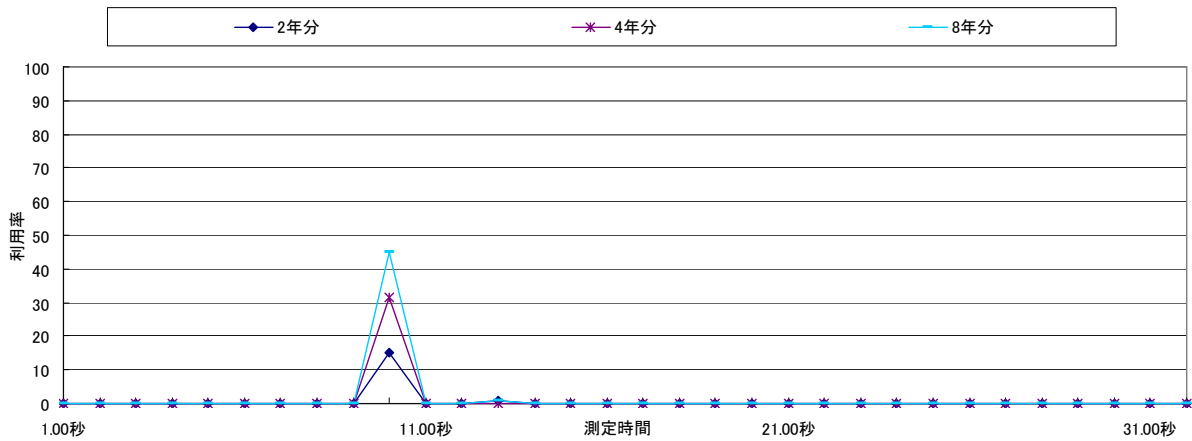


図 6-3-2-8. 文献数増加時の DB サーバの CPU 利用率
(1 コアあたりの CPU 利用率)

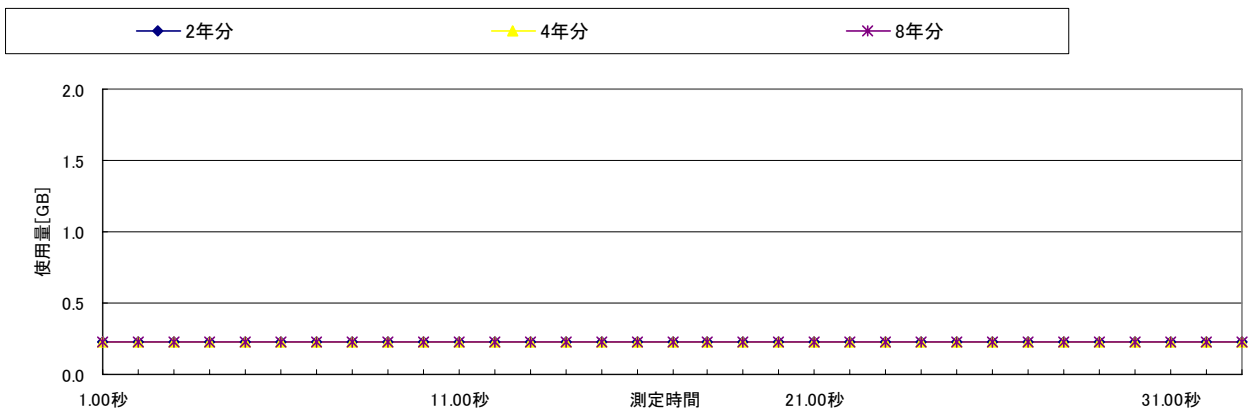


図 6-3-2-9. 文献数増加時の DB サーバのメモリ利用量

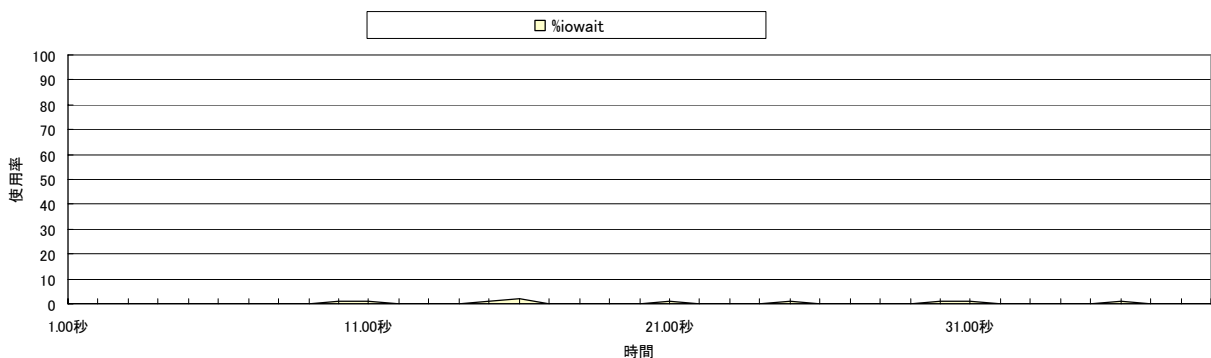


図 6-3-2-10. DB サーバの I/Owait 利用率 (8 年分)

(c) AP サーバ(分散サーバ数増加時)

分散サーバ数増加時の CPU 利用率を図 6-3-2-1 1 に示す。図は AP サーバの 4 コアのうち、実際に CPU に負荷がかっていた 1 コアのみでのグラフである。分散サーバ数に依存せず、AP サーバの CPU 利用率が高いことが分かる。また、メモリ利用量はほぼ一定となる(図 6-3-2-1 2)。なお、いずれの検証パターンにおいても、I/Owait は発生していない(図 6-3-2-1 3 参照)。

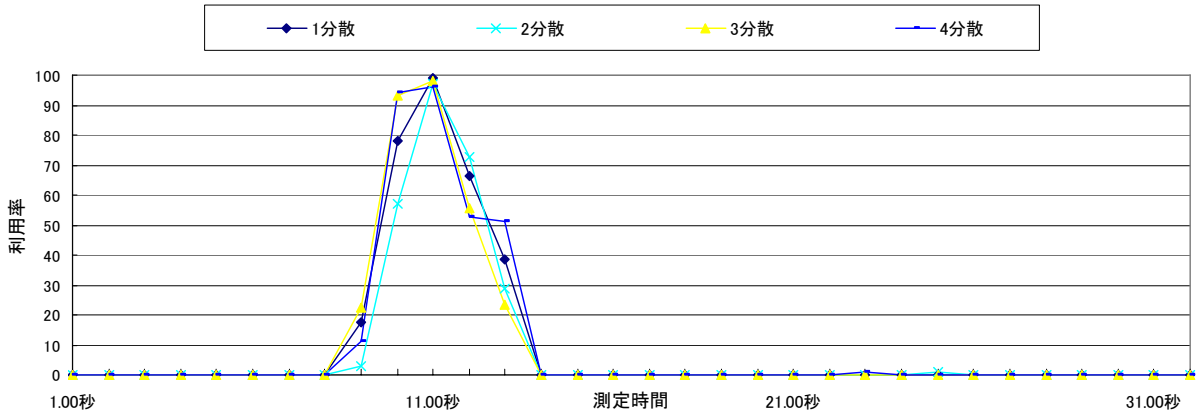


図 6-3-2-1 1. 分散サーバ数増加時の AP サーバの CPU 利用率 (1 コアあたりの CPU 利用率)

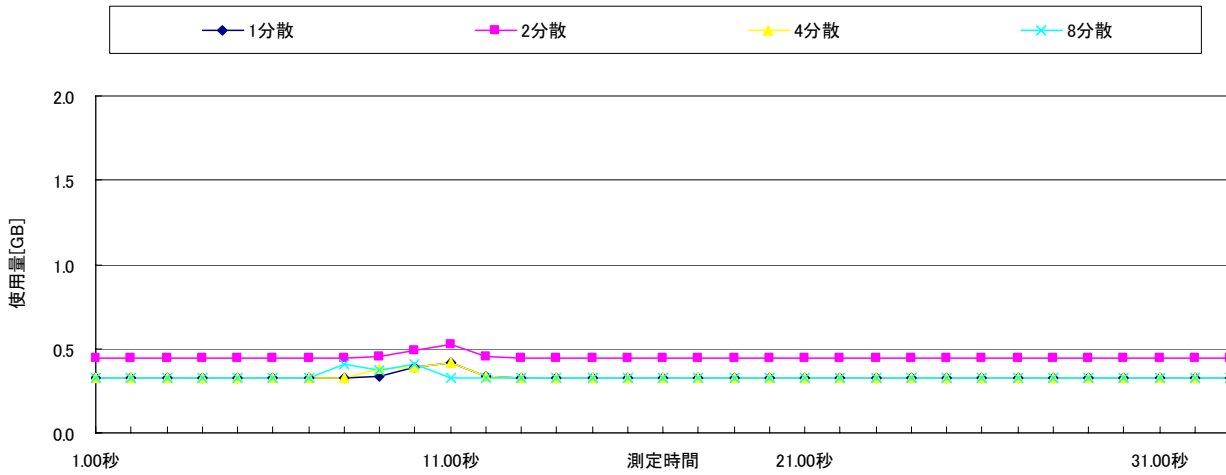


図 6-3-2-1 2. 分散サーバ数増加時の AP サーバのメモリ利用量

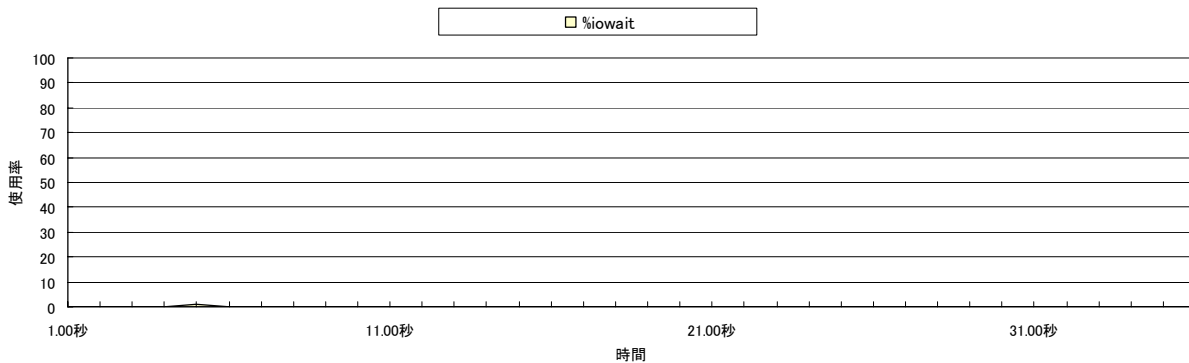


図 6-3-2-1 3. AP サーバの I/Owait 利用率 (8 分散)

(d) DB サーバ(分散サーバ数増加時)

分散サーバ数増加時の DB サーバの CPU 利用率を図 6-3-2-14 に示す。また、メモリ使用状況を図 6-3-2-15 に示す。分散サーバ数に比例して、CPU1 コアあたりの CPU 利用率が減少していることが分かる。また、メモリ利用量はほぼ一定となる。なお、いずれの検証パターンにおいても、WAM はメモリ上に展開され、I/Owait は発生していない(図 6-3-2-16 参照)。

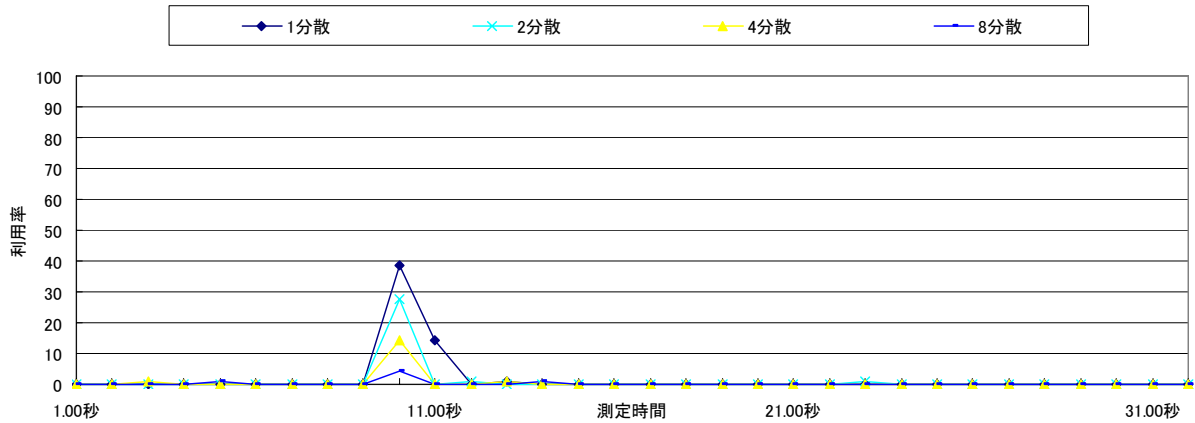


図 6-3-2-14. 分散サーバ数増加時の DB サーバの CPU 利用率
(1 コアあたりの CPU 利用率)

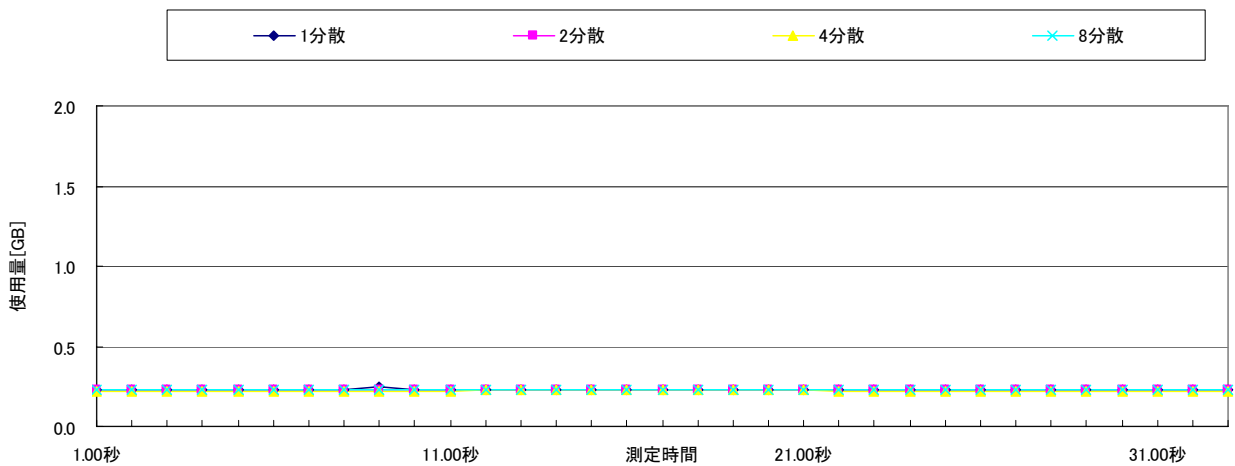


図 6-3-2-15. 分散サーバ数増加時の DB サーバのメモリ利用量

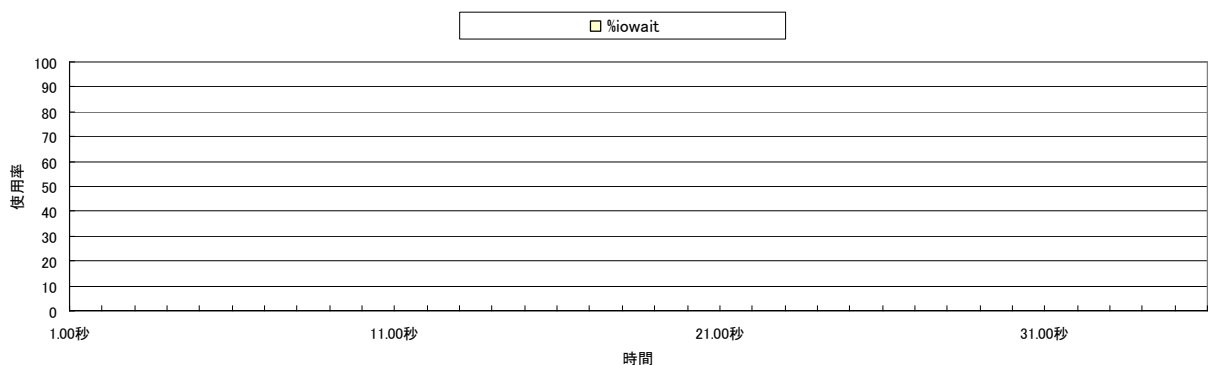


図 6-3-2-16. DB サーバの I/Owait 利用率 (8 分散)

(4) クライアントPCリソース使用状況

クライアントPCのリソース使用状況を図6-3-2-17に示す。データマイニングでは、概念検索と比べてPCのCPU利用率が高い傾向にある。これは、グラフ表示の中のスペクトル表示について、クライアントPCに負荷がかかるためである。

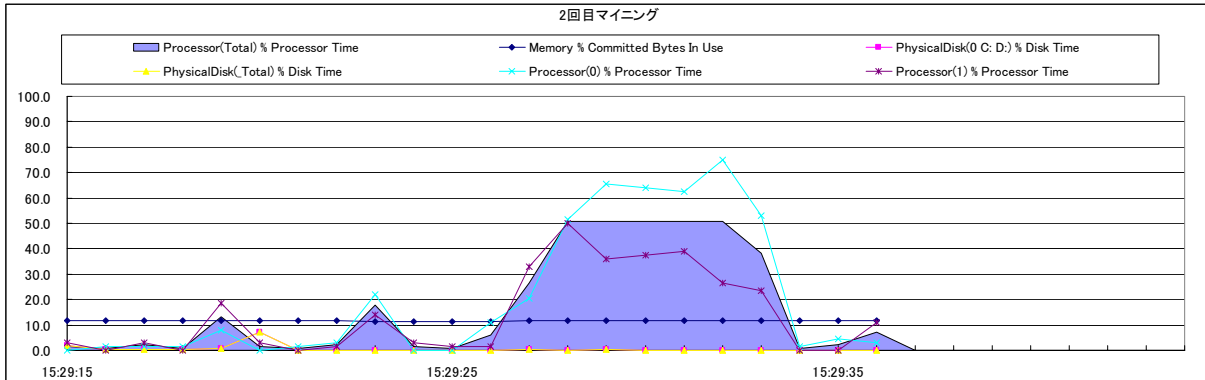


図6-3-2-17. クライアントPCのリソース利用率

6-3-3. 測定結果のまとめ

データマイニングの測定結果を表6-3-3-1にまとめる。

表6-3-3-1. データマイニングの測定結果

#	マシン	測定結果	参照先
1	AP サーバ	分散サーバ数に依存せず、AP サーバの処理時間はほぼ一定である	図6-3-2-2
2		文献数の増加に応じて、スペクトル表示のための特徴語数が増加するため、処理時間は伸びる	図6-3-2-1
3		CPU は 1 コアのみ負荷がかかるが、概念検索と比較してその利用率は高い	表6-3-2-1 表6-3-2-2
4		CPU を利用するのは、GETA による検索よりも、UP の処理自体である	6-3-2 章
5		I/Owait やメモリなどのボトルネックは発生しない	図6-3-2-6 図6-3-2-7 図6-3-2-12 図6-3-2-13
6	DB サーバ	文献数に比例し、DB サーバの総 CPU 時間は増加する	図6-3-2-1
7		分散サーバ数に比例して、負荷がかかる CPU コア数は増加する	表6-3-2-2
8		分散サーバ数が増加すると、CPU1 コアあたりの CPU 時間は減少する	表6-3-2-2
9		メモリのボトルネックは発生しない	図6-3-2-9 図6-3-2-15
10		WAM はメモリ上に展開され、I/Owait のボトルネックは発生しない	図6-3-2-10 図6-3-2-16
11	クライアント PC	分散サーバ数に依存せず、クライアント PC の利用率はほぼ一定である	図6-3-2-17
12		文献数の増加に応じて、スペクトル表示のための特徴語数が増加するため、処理時間は伸びる	図6-3-2-1
13		メモリ利用量は小さいが、グラフ表示に関する CPU 利用率が高く、本調査で用いた UP では、目標性能を満たせない可能性がある	6-3-2 章 図6-3-2-17

6-3-4. 処理時間に対する改善策

測定結果より、本調査で用いたデータマイニング機能では目標性能を満たせない可能性がある。本調査で用いたデータマイニング機能はDBサーバに対する検索時間は短いものの、APサーバ、PCともにグラフ表示に関するUPの処理時間が大きい。これは今回の調査にあたり、機能追加を行ったスペクトル表示機能による影響である。スペクトル表示の追加前は、実際にグラフ表示エリアに表示される項目数(デフォルトは35個)分だけ、PCやAPサーバでは特徴語(ワードや分類)の処理を行っていた。今回、機能追加を行ったスペクトル表示では、対象の文献集合が持つすべての特徴語を処理するため、PC、APサーバともに処理時間が増加する。

実際にスペクトル表示を行う際に、すべての特徴語を同時に表示するわけではないため、この処理を改善することで、概念検索と同程度まで処理時間の短縮が可能である。

表6-3-4-1. 特徴語数と処理時間

#	項目	概念検索	データマイニング	データマイニング (スペクトルあり)
1	特徴語数	30件	35件	数千件~数万件
2	APサーバ処理時間	約0.15秒	約0.18秒	約3.30秒
3	PCクライアント時間	約0.40秒	約0.47秒	約5.91秒

6-4. 基礎数値の算出

(1) 算出方法

概念検索では文献数、分散サーバ数に依存せず、PC、APサーバの処理時間はほぼ一定である。データマイニングについては、分散サーバ数に依存せず、PC、APサーバの処理時間はほぼ一定である。データマイニングでは、スペクトル表示の特徴語数、すなわち文献数によりPCとAPサーバの処理時間は増加する。データマイニングの実際の利用シーンを想定した場合、特徴語全件の取得は現実的ではないため、ここでは特徴語の処理量は一定であるとし、モデルとしては2年分の文献数を想定する(実際の利用にあたっては、処理する特徴語量はさらに少なくなる)。

また、概念検索、データマイニングともにDBサーバについては、以下の傾向が挙げられる。

- ・ 文献数に比例して、DBサーバのCPU時間は増加する
- ・ 分散サーバ数に比例して、DBサーバのCPU時間は減少する

以上の傾向から、単位文献数あたりのDBサーバのCPU時間を求め、目標性能を満たすために必要な分散サーバ数とその時の1分散あたりの文献数を算出すればよい。

ここでは以下の方法で単位文献数(1000件あたり)のDBサーバのCPU処理時間を算出する。

$$[\text{単位文献数あたりの総CPU時間}] = [\text{DBサーバの総CPU時間}] \div [\text{文献数}]$$

(2) 概念検索の基礎数値

検証パターン1から3までの、文献数と総CPU時間より求めた1000件あたりの総CPU時間を表6-4-1に示す。3つの検証パターンから最大のものを利用する。なお、ここで約25%の安全率を見込み、基礎数値を0.008秒とする。

表6-4-1. 概念検索の単位件数あたりの総CPU時間

パターン	年数	文献数 (件)	総CPU時間 (秒)	1000件換算時の 総CPU時間(秒)
検証パターン1	2年分	687,881	4.277	0.00622
検証パターン2	4年分	1,359,857	8.690	0.00639
検証パターン3	8年分	2,774,380	17.720	0.00639

(3) データマイニングの基礎数値

検証パターン1から3までの、文献数と総CPU時間より求めた1000件あたりの総CPU時間を表6-4-2に示す。3つの検証パターンから最大のものを利用する。なお、ここで約25%の安全率を見込み、基礎数値を0.0015秒とする。

表6-4-2. データマイニングの単位件数あたりの総CPU時間

パターン	年数	文献数 (件)	総CPU時間 (秒)	1000件換算時の 総CPU時間(秒)
検証パターン1	2年分	615,932	0.713	0.00116
検証パターン2	4年分	1,238,343	1.441	0.00116
検証パターン3	8年分	2,519,170	2.010	0.00080

6-5. 本番環境を想定したサイジング

ここでは、概念検索とデータマイニングを本番環境に適用した場合、検索時間の目標値である 5.0 秒を満たすために必要なサーバ台数を算出する。なお、ここでいう検索時間の目標値はクライアント PC の応答時間であり、検索を実行してから結果が表示されるまでの時間である。

データマイニングについてはUPの仕様や構成によって処理時間が大きく異なることからここでは、あくまでも参考値としての扱う。

なお、サイジングにあたって前提とする業務量(トランザクション量)は、あくまでも想定値である。次世代検索システムにおける前提業務量はまだ定義されていないため、今後の最適化計画の実行方針に従って適宜見直す必要がある。

(1) 想定条件

(a) 概念検索

サイジングを行うにあたって本番環境を表 6-5-1 のとおり想定する。概念検索の検索対象文献数は、2008 年 12 月時点の件数とする。また、トランザクション量は、現行の全文検索のピーク時と同等とする。また、サイジングで想定するサーバは、本性能測定で使用したサーバとする。

表 6-5-1. サイジング時の前提環境

#	適用箇所	項目	内容	前提条件の理由
1	共通	検索対象文献数	1600 万件	2008 年 12 月時点での国内公報・非特許テキスト件数(案件単位)
2	庁内	目標検索時間	5 秒	現行の全文検索の目標値
3		トランザクション量	8.7TPS	現行の本番の全文検索のピーク時のトランザクション量
4	庁外	目標検索時間	8 秒	現行の IPDL のレスポンス
5		トランザクション量	2.5TPS	IPDL のトランザクション量を想定

(b) データマイニング

サイジングを行うにあたって本番環境を表 6-5-2 のとおり想定する。データマイニングの検索対象文献数は、2008 年 12 月時点の件数とする。また、トランザクション量は、現行の検索キー照会と同等とする。また、サイジングで想定するサーバは、本性能測定で使用したサーバとする。

性能目標値については、本来庁内 5 秒、庁外 8 秒であるが、今回の検証ツールでは UP の特性上、目標性能を満たせないことから、10 秒を仮の目標値としてサイジングを行う。

表 6-5-2. サイジング時の前提環境

#	適用箇所	項目	内容	前提条件の理由
1	共通	検索対象文献数	1600 万件	2008 年 12 月時点での国内公報・非特許文献蓄積件数(案件数)
2	庁内	目標検索時間	10 秒 (5 秒)	庁内の検索における目標値は 5 秒
3		トランザクション量	0.694TPS	現行の検索キー照会のトランザクション量より推定
4	庁外	目標検索時間	10 秒 (8 秒)	庁外の検索における目標値は 8 秒
5		トランザクション量	0.313TPS	現在、データマイニングの対外提供は行われていないため、前提業務量の指針がない。このため、1 時間あたり 1000 回利用される想定とした

(2) サイジングの考え方

(a) 負荷発生時の処理時間の算出方法

測定結果より以下の条件で負荷発生時の処理時間を机上算出し、サーバの必要数を求める

- ・ トランザクションの負荷分散により n 台の AP サーバに均等に分散される
- ・ AP サーバ 1 台あたりの DB サーバは m 台 (DB サーバの総台数は $m \times n$)
- ・ AP サーバにて待ち行列が発生する $M/M/1$ モデル
- ・ AP サーバ 1 台あたりのトランザクションは $1/n$ となる

以降で使用する処理時間について、待ち行列を考慮した処理時間を図 6-5-1、表 6-5-3、表 6-5-4 のとおり定義し、性能測定の結果より得られる単件処理の処理時間を図 6-5-2、表 6-5-5 のとおり定義する。

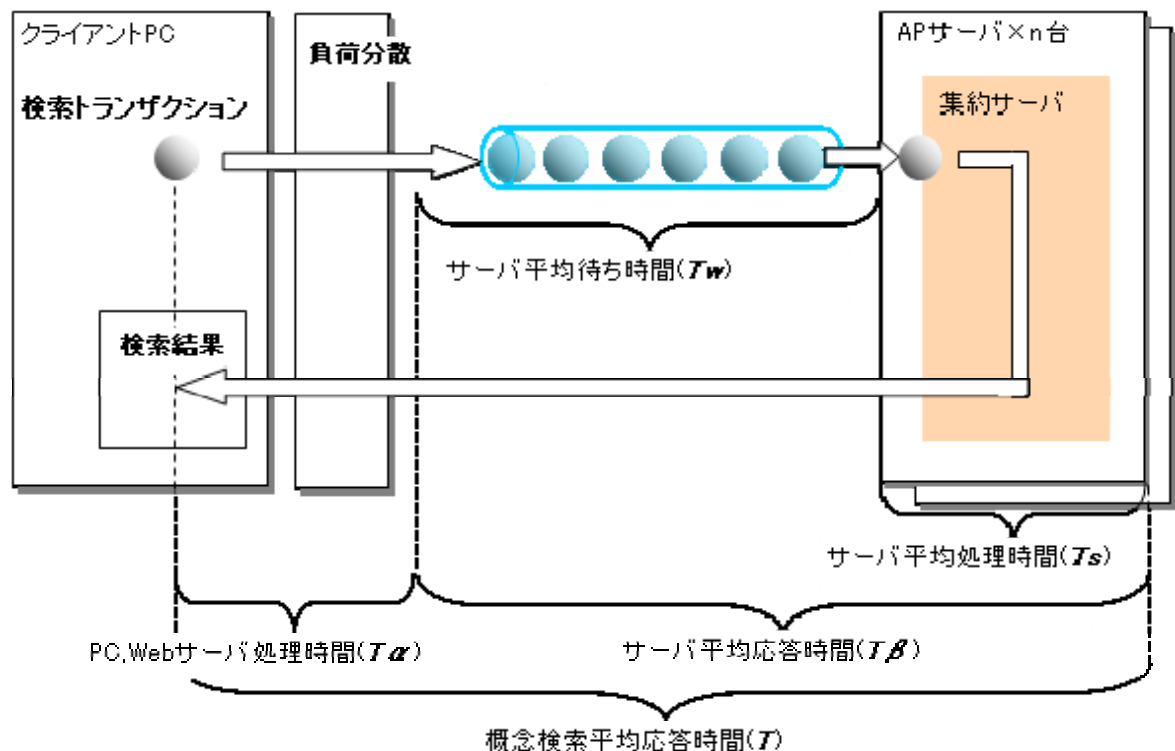


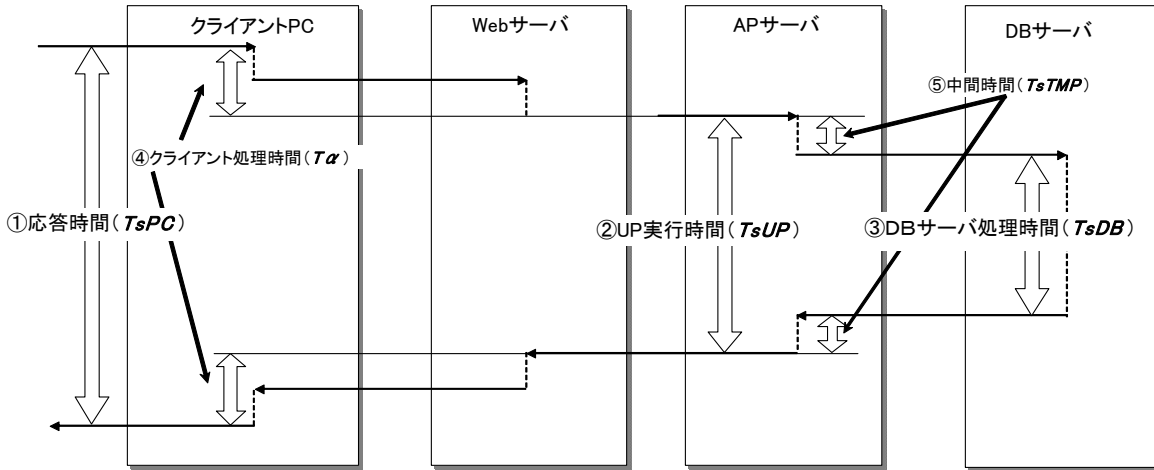
図 6-5-1. 概念検索の待ち行列理論を考慮した平均応答時間イメージ

表 6-5-3. 各処理時間の説明

#	項目	説明	算出式	備考
1	T	概念検索平均応答時間	$T = T_{\alpha} + T_{\beta}$	
2	T_{α}	PC、Webサーバ処理時間	-	図 6-5-2 参照
3	T_{β}	サーバ平均応答時間	$T_{\beta} = T_s + T_w$	APサーバとDBサーバの応答時間の合計値
4	T_w	サーバ平均待ち時間	$T_w = \rho / (1 - \rho) \times T_s$	<ul style="list-style-type: none"> ・ $M/M/1$ モデルの待ち行列理論から算出。ρ は、表 6-5-4 を参照 ・ APサーバとDBサーバの応答時間の合計値
5	T_s	サーバ平均処理時間	基礎数値を使用する	APサーバとDBサーバの応答時間の合計値

表 6-5-4. 待ち行列の計算で使用する項目

#	項目	名称	説明
1	λ	平均到着率	単位時間あたりに到着するトランザクション数
2	μ	平均サービス率	単位時間あたりに処理できる件数
3	ρ	平均利用率	単位時間あたりに処理できる件数に対して、到着するトランザクション数の割合 ($\rho = \lambda / \mu$)



※ネットワーク上の処理時間は、PC・サーバ内の時間に含まれている。

図 6-5-2. 単件処理の場合の応答時間イメージ

表 6-5-5. 各処理時間の説明

#	項目	説明
1	TsPC	応答時間
2	Tα	クライアント PC の応答時間から AP サーバ内の時間を引いた時間。 実質的には、クライアント PC と Web サーバで費やした時間
3	TsTMP	AP サーバ内の時間から DB サーバ処理時間を引いた時間。 実質的には、AP サーバで費やした時間。
4	TsUP	UP 実行時間。AP サーバと DB サーバで費やした時間。
5	TsDB	DB サーバ処理時間。DB サーバで費やした時間

(b) サイジング方法

今回の性能測定結果より、DB サーバ側の総 CPU 処理時間は文献数に比例して増加し、1分散サーバあたりの CPU 処理時間は分散サーバ数に比例して減少することが分かっている。そのため、まず、1分散サーバで受け持つ文献数の限界を調査する。そこで、1分散サーバで処理する事を前提にして、検索対象文献数を1000件ずつ増加させ、文献数ごとの平均応答時間を算出する。当然、待ち行列の理論上、想定量のトランザクションが発生し続けるため、文献数がある一定数に達すると、目標値を満たせなくなる時点がある。その時の文献数をもとにして、(1)で想定した検索対象文献数1600万件においてはいくつの分散数が必要かを求め、サーバ台数を算出する。

(c) サイジングの計算式

概念検索の平均応答時間を目標値以内にする必要がある。概念検索の平均応答時間は、図6-5-1、図6-5-2、表6-5-3、表6-5-4、表6-5-5で示したとおり、以下の式で表すことができる。

$$T = T\alpha + T_s + T_w$$

$$T_s = T_{sDB} + T_{sTMP}$$

ここで、 $T\alpha$ と T_{sTMP} は、性能測定結果から取得できる。表6-5-6よりほぼ一定の時間となっているため、サイジング時にはこの平均値を使用する。また、 T_w は、表6-5-3、表6-5-4より、待ち行列理論(M/M/1モデル)から算出し、 T_{sDB} に関しては、6-4章の基礎数値より、文献数1000件あたりの処理時間0.008秒を使用する(表中の T_{sDB} は T_{sTMP} を求めめるために使用する)。また、データマイニングの測定結果を表6-5-7に示す。データマイニングでは文献数2年分のパターンをモデルとし、文献数1000件あたりの処理時間を0.0015秒とする。

表6-5-6. 性能測定結果 (概念検索)

処理フェーズ	性能測定パターンごとの処理時間 (秒)							平均時間 (秒)
	パターン1	パターン2	パターン3	パターン4	パターン5	パターン6	パターン7	
①TsPC	1.578	2.641	5.036	4.896	2.615	1.599	1.052	—
②TsUP	1.465	2.531	4.902	4.609	2.495	1.424	0.943	—
③TsDB	1.085	2.150	4.500	4.230	2.092	1.055	0.540	—
①-② ($T\alpha$)	0.113	0.110	0.134	0.287	0.120	0.175	0.109	0.150
②-③ (T_{sTMP})	0.380	0.381	0.402	0.379	0.403	0.369	0.403	0.388

表6-5-7. 性能測定結果 (データマイニング)

処理フェーズ	性能測定パターンごとの処理時間 (秒)					平均時間 (秒)
	パターン1	パターン4	パターン5	パターン6	パターン7	
①TsPC	8.927	9.229	8.928	9.354	9.630	9.214
②TsUP	3.465	3.947	3.115	3.287	3.584	3.480
③TsDB	0.713	0.687	0.599	0.670	0.477	0.630
①-② ($T\alpha$)	5.462	5.282	5.813	6.067	6.046	5.734
②-③ (T_{sTMP})	2.752	3.260	2.516	2.617	3.107	2.851

(3) 庁内目標値に対するサイジング（概念検索）

単位件数あたりの DB サーバ側の総 CPU 処理時間(0.008 秒)と待ち行列モデルより求めた、検索対象文献数ごとの平均応答時間を表 6-5-8 に示す。計算結果より、AP サーバが 16 台の時が最小構成となるため、ここでは AP サーバ 16 台の構成を示す。庁内の目標値 5 秒を満たすためには、1 分散サーバあたりの文献数が 11.8 万件以内である必要がある。これを、本番環境と同じ 1600 万件分処理するには、 $\uparrow 1600 \text{ 万件} / 11.8 \text{ 万件} \uparrow = 136$ の数だけ分散サーバが必要である。

表 6-5-8. 検索対象文献数ごとの概念検索応答時間（庁内）（APサーバ16台）

#	検索対象文献数 (万件)	平均応答時間(秒)	目標値判定	備考
		1分散サーバ	庁内	
1	1.0	0.778	○	
2	2.0	0.931	○	
3	3.0	1.104	○	
4	4.0	1.301	○	
5	5.0	1.529	○	
6	11.5	4.679	○	
7	11.6	4.776	○	
8	11.7	4.877	○	
9	11.8	4.980	○	
10	11.9	5.087	×	1分散サーバの処理時間が到着するトランザクション量に対応できない

今回の検証環境の場合、サーバのコア数が 4 であるためサーバ 1 台あたり最大 4 つの分散サーバを使用できる。従って、目標値 5 秒を満たすための必要なサーバ台数は以下のとおりとなる。

【APサーバ】 16 コア / 4 コア = 4 台

【DBサーバ】 136 / 4 コア = 34 台（総 DB サーバ数 16 × 34 = 544 台）

(4) 庁外目標値に対するサイジング（概念検索）

単位件数あたりの DB サーバ側の総 CPU 処理時間(0.008 秒)と待ち行列モデルより求めた、検索対象文献数ごとの平均応答時間を表 6-5-9 に示す。計算結果より、AP サーバが 6 台の時が最小構成となるため、ここでは AP サーバ 6 台の構成を示す。庁外の目標値 8 秒を満たすためには、1 分散サーバあたりの文献数が 18.0 万件以内である必要がある。これを、本番環境と同じ 1600 万件分処理するには、 $\uparrow 1600 \text{ 万件} / 18 \text{ 万件} \uparrow = 89$ の数だけ分散サーバが必要である。

表 6-5-9. 検索対象文献数ごとの概念検索応答時間（庁外）（APサーバ6台）

#	検索対象文献数 (万件)	平均応答 時間(秒)	目標 値判 定	備考
		1分散サ ーバ	庁内	
1	1.0	0.731	○	
2	2.0	0.860	○	
3	3.0	1.000	○	
4	4.0	1.154	○	
5	5.0	1.323	○	
6	15.0	4.843	○	
7	16.0	5.618	○	
8	17.0	6.584	○	
9	18.0	7.819	○	
10	19.0	9.457	×	1 分散サーバの処理時間が到着する トランザクション量に対応できない

今回の検証環境の場合、サーバのコア数が 4 であるためサーバ 1 台あたり最大 4 つの分散サーバを使用できる。従って、目標値 5 秒を満たすための必要なサーバ台数は以下のとおりとなる。

【APサーバ】 $\uparrow 6 \text{ コア} / 4 \text{ コア} \uparrow = 2$ 台

【DBサーバ】 $\uparrow 89 / 4 \text{ コア} \uparrow = 23$ 台（総 DB サーバ数 $6 \times 23 = 138$ 台）

(5) 庁内目標に対するサイジング（データマイニング）（参考）

単位件数あたりの DB サーバ側の総 CPU 処理時間(0.0015 秒)と待ち行列モデルより求めた、検索対象文献数ごとの平均応答時間を表 6-5-10 に示す。計算結果より、AP サーバが 12 台の構成となるため、ここでは AP サーバ 12 台の構成を示す。庁内の目標値 10 秒を満たすためには、1 分散サーバあたりの文献数が 38.0 万件以内である必要がある。これを、本番環境と同じ 1600 万件分処理するには、 $\uparrow 1600 \text{ 万件} / 38 \text{ 万件} \uparrow = 43$ の数だけ分散サーバが必要である。

表 6-5-10. 検索対象文献数ごとのデータマイニング応答時間（庁内）（APサーバ12台）

#	検索対象文献数 (万件)	平均応答時間(秒)	目標値判定	備考
		1分散サーバ	庁内	
1	1.0	9.169	○	
2	2.0	9.190	○	
3	3.0	9.212	○	
4	4.0	9.234	○	
5	5.0	9.255	○	
6	10.0	9.365	○	
7	20.0	9.587	○	
8	30.0	9.813	○	
9	38.0	9.998	○	
10	39.0	10.022	×	1分散サーバの処理時間が到着するトランザクション量に対応できない

今回の検証環境の場合、サーバのコア数が 4 であるためサーバ 1 台あたり最大 4 つの分散サーバを使用できる。従って、目標値 10 秒を満たすための必要なサーバ台数は以下のとおりとなる。

【APサーバ】 12 コア / 4 コア = 3 台

【DBサーバ】 $\uparrow 43 / 4 \text{ コア} \uparrow = 11$ 台（総 DB サーバ数 $12 \times 11 = 132$ 台）

(6) 庁外目標値に対するサイジング（データマイニング）（参考）

単位件数あたりの DB サーバ側の総 CPU 処理時間(0.0015 秒)と待ち行列モデルより求めた、検索対象文献数ごとの平均応答時間を表 6-5-11 に示す。計算結果より、AP サーバが 6 台の時の最小構成となるため、ここでは AP サーバ 6 台の構成を示す。庁内の目標値 10 秒を満たすためには、1 分散サーバあたりの文献数が 42.0 万件以内である必要がある。これを、本番環境と同じ 1600 万件分処理するには、 $\uparrow 1600 \text{ 万件} / 42 \text{ 万件} \uparrow = 39$ の数だけ分散サーバが必要である。

表 6-5-11. 検索対象文献数ごとのデータマイニング応答時間（庁外）（APサーバ6台）

#	検索対象文献数 (万件)	平均応答 時間(秒)	目標 値判 定	備考
		1分散サ ーバ	庁内	
1	1.0	9.103	○	
2	2.0	9.124	○	
3	3.0	9.145	○	
4	4.0	9.166	○	
5	5.0	9.187	○	
6	10.0	9.292	○	
7	20.0	9.504	○	
8	30.0	9.721	○	
9	40.0	9.942	○	
10	42.0	9.987	○	
11	43.0	10.009	×	1 分散サーバの処理時間が到着する トランザクション量に対応できない

今回の検証環境の場合、サーバのコア数が 4 であるためサーバ 1 台あたり最大 4 つの分散サーバを使用できる。従って、目標値 10 秒を満たすための必要なサーバ台数は以下のとおりとなる。

【APサーバ】 $\uparrow 6 \text{ コア} / 4 \text{ コア} \uparrow = 2$ 台

【DBサーバ】 $\uparrow 39 / 4 \text{ コア} \uparrow = 10$ 台（総 DB サーバ数 $6 \times 10 = 60$ 台）

6-6. データ蓄積

6-6-1. 蓄積データ概要

汎用連想計算エンジン GETA で検索するためのデータである WAM を作成するには、要素と要素の出現頻度の格納された頻度ファイルと要素の様々な情報が格納された補助ファイルが必要となる。

図 6-6-1-1 で示すように、一次情報よりバッチ処理により頻度ファイル及び補助ファイルを作成し、頻度ファイル及び補助ファイルから GETA 標準の WAM 作成ユーティリティにより WAM を作成する。

本節では頻度ファイル、補助ファイル、WAM の蓄積の性能について検証する。

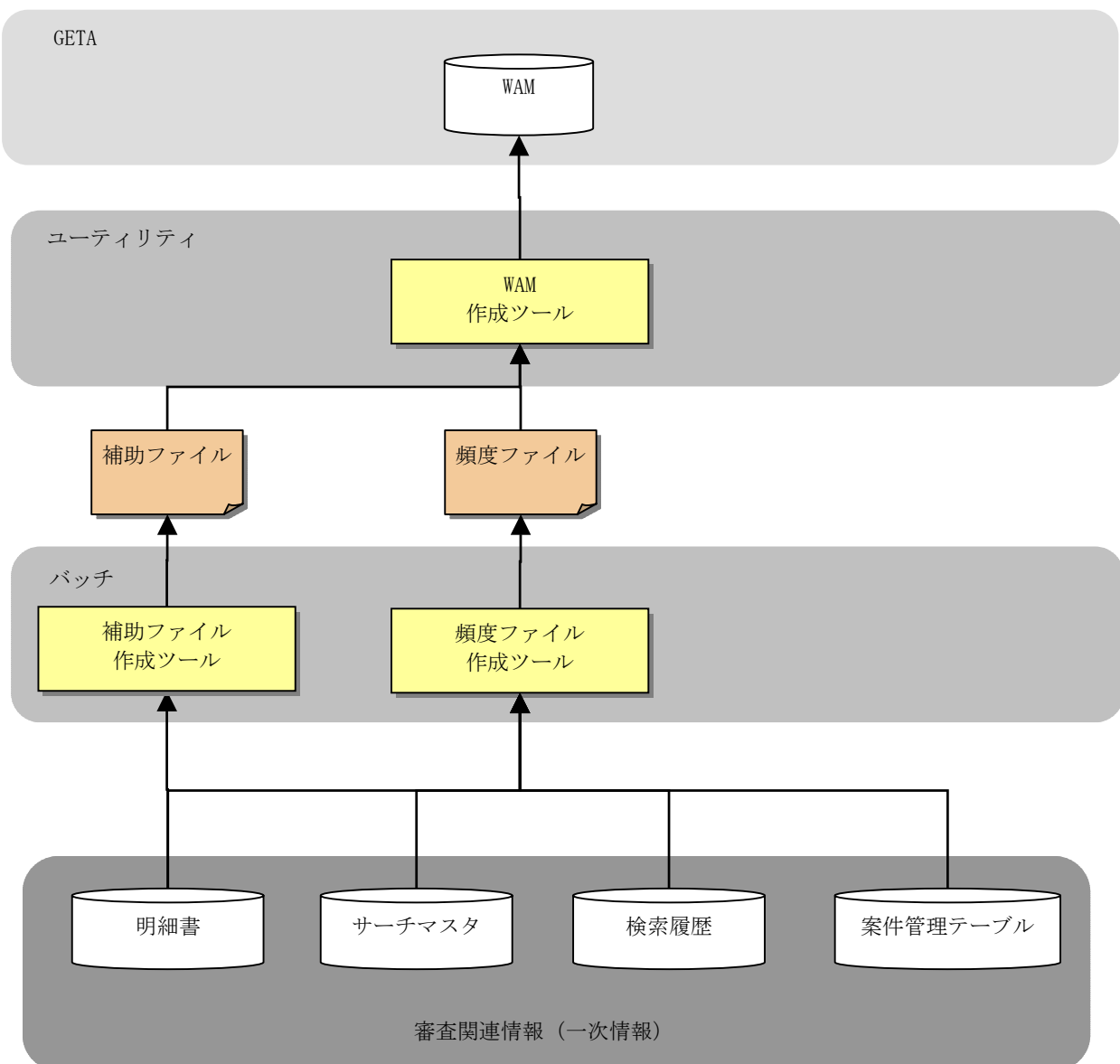


図 6-6-1-1. WAM 作成手法

6-6-2. 蓄積データ

検証対象とする蓄積データを表6-6-2-1に示す。

表6-6-2-1. 検証対象となる蓄積データ一覧

#	ファイル名称	一次情報 ^{*1}	データ内容
1	公報単語 頻度ファイル	明細書	明細書から文献ごとのワードと出現頻度を抽出して格納したファイル。
2	概念検索 頻度ファイル	明細書 サーチマスタ 案件管理マスタ	明細書から文献ごとのワードと出現頻度抽出し、サーチマスタの公開基準日・公知日・テーマ・FI・Fターム、案件管理マスタの主テーマを格納したファイル。
3	概念検索特定箇所N倍 頻度ファイル	明細書 サーチマスタ 案件管理テーブル	明細書から文献ごとのワードと出現頻度を抽出して明細書の特定箇所(要約・請求項・請求項1・実施例)の出現頻度をN倍し、サーチマスタの公開基準日・公知日・テーマ・FI・Fターム、案件管理マスタの主テーマを格納したファイル。
4	共起関連分類 頻度ファイル	サーチマスタ 明細書	サーチマスタから文献ごとのテーマ・FI・Fタームを格納したファイル。 ※明細書中の分類が古い可能性があるためサーチマスタより最新の分類を抽出
5	関連発明者 頻度ファイル	明細書	明細書から発明者・出願人を抽出して格納したファイル。
6	本願別 検索式履歴ファイル	検索履歴 (検索式履歴)	検索履歴から本願番号ごとのテーマ・検索式履歴を格納したファイル。
7	印刷履歴活用本願別 検索式履歴ファイル	検索履歴 (検索式履歴) (スクリーニング)	検索履歴から検索でヒットした文献が1回以上印刷されたことのある本願番号ごとのテーマ・検索式履歴を格納したファイル。
8	関連検索キー 頻度ファイル	検索履歴 (検索式履歴) (スクリーニング)	本願別検索式履歴ファイル(#6、#7)から検索キーを抽出して格納したファイル。
9	検索式履歴ワード 頻度ファイル	検索履歴 (検索式履歴) (スクリーニング)	本願別検索式履歴ファイル(#6、#7)から検索キー(全文(/TX))を抽出して格納したファイル。
10	検索式履歴分類 頻度ファイル	検索履歴 (検索式履歴) (スクリーニング)	本願別検索式履歴ファイル(#6、#7)から検索キー(分類(テーマ・/FI・/FT))を抽出して格納したファイル。
11	審査関連情報 頻度ファイル	明細書 サーチマスタ 検索履歴 (検索式履歴) (スクリーニング)	明細書から文献ごとのワードと出現頻度・発明者・出願人を抽出し、サーチマスタのテーマ・FI・Fターム、本願別検索式履歴ファイル(#6、#7)から検索キーを抽出して格納したファイル。 ※明細書中の分類が古い可能性があるためサーチマスタより最新の分類を抽出
12	補助ファイル	明細書	明細書から発明の名称を抽出して格納したファイル。
13	WAM	頻度ファイル	GETA標準のWAM作成ユーティリティにより作成されるGETAで検索可能なデータ。

*1: WAM生成のために情報抽出元としたデータベース

(1) 頻度ファイル

要素と要素の出現頻度を格納するファイル。

頻度ファイルは、頻度ファイル作成ツールにおいて作成単位の文献集合を任意に指定できることから、並列処理で頻度ファイルを作成できる。そのため、作成時間を短縮することが可能である。また、1文献に対し1ファイルまたは複数文献に対し1ファイルという単位で作成が可能であるため、一次情報に追加、変更があった場合、該当する1ファイルを更新するだけでよく、特定の頻度ファイルのみを追加することができる。並列処理、頻度ファイルの追加についてのイメージ図を図6-6-2-1に示す。

(2) 補助ファイル

頻度ファイルの要素の多様な情報を格納するファイル。

補助ファイルも頻度ファイルと同様に、補助ファイル作成ツールにおいて作成単位の文献集合を任意に指定できることから、並列処理で補助ファイルを作成できる。そのため、作成時間を短縮することが可能である。また、1 文献に対し 1 ファイルまたは複数文献に対し 1 ファイルという単位で作成が可能であるため、一次情報に追加、変更があった場合、該当する 1 ファイルを更新するだけでよく、特定の補助ファイルのみを追加することができる。

(3) WAM

要素と要素の出現頻度は頻度ファイルから抽出し、結果表示時の付加情報を補助ファイルから抽出して GETA で検索できるようにしたデータ。

現行の最新バージョンの GETA では差分更新がサポートされていないため、頻度ファイルに追加、変更があった場合には該当する WAM を再作成する必要がある。

ただし技術的には可能であるため将来的にはサポートされる可能性はある。

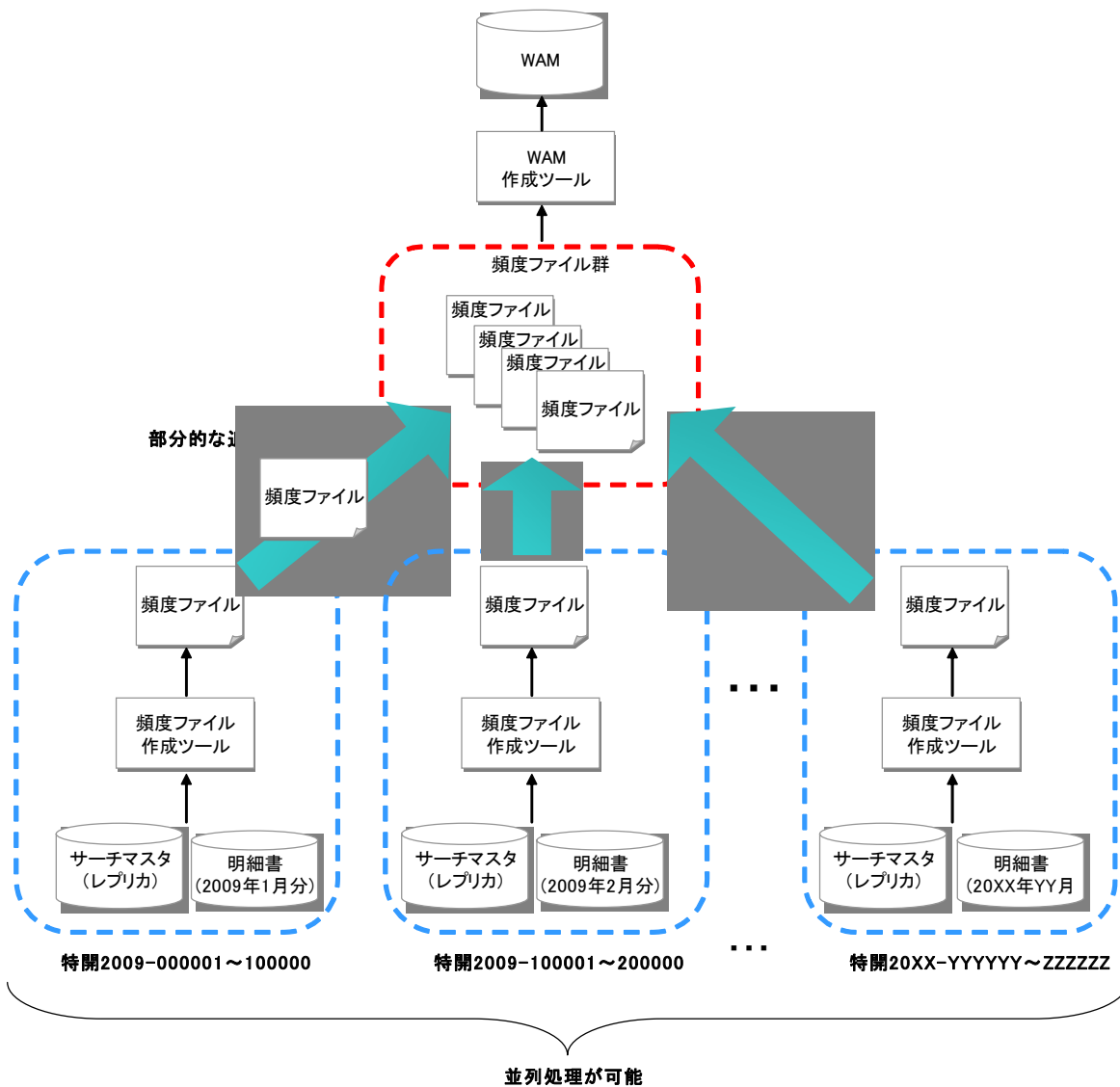


図 6-6-2-1. 頻度ファイル並列処理、追加作成例

6-6-3. 測定方法

(1) 頻度ファイル、補助ファイル蓄積処理の測定

測定環境は AP サーバを使用する。

(a) 測定方法

入力データを 1 万件に設定し、処理時間、ファイルサイズを測定する。

測定中はサーバリソースの使用状況を監視する。

(b) 全件データでのファイルサイズ及び蓄積処理性能の算出

測定結果からファイルサイズ単価を算出する。

$$\text{[ファイルサイズ単価]} (\text{byte/件}) = \text{[ファイルサイズ]} (\text{byte}) \div 10,000 (\text{件})$$

ファイルサイズ単価から全件データでのファイルサイズを算出する。

$$\text{[全件ファイルサイズ]} (\text{byte}) = \text{[ファイルサイズ単価]} (\text{byte/件}) \times \text{[全件データ]} (\text{件})$$

測定結果から性能単価を算出する。

$$\text{[性能単価]} (\text{秒/件}) = \text{[処理時間]} (\text{秒}) \div 10,000 (\text{件})$$

単価性能から全件データでの蓄積処理時間を算出する。

$$\text{[全件蓄積処理性能]} (\text{秒}) = \text{[性能単価]} (\text{秒/件}) \times \text{[全件データ]} (\text{件})$$

(c) サーバリソース使用状況の取得

CPU 使用率、I/O wait 発生率、メモリ使用率をサーバリソース使用状況から取得する。

(2) WAM 蓄積処理の測定

測定環境は AP サーバ・DB サーバ 1・DB サーバ 2 を使用する。

(a) 測定方法

入力データを 1 万件に設定し、ファイルサイズ、処理時間を測定する。

測定中はサーバリソースの使用状況を監視する。

(b) 全件データでのファイルサイズ及び蓄積処理性能の算出

測定結果からファイルサイズ単価を算出する。

$$\text{[ファイルサイズ単価]} (\text{byte/件}) = \text{[ファイルサイズ]} (\text{byte}) \div 10,000 (\text{件})$$

ファイルサイズ単価から全件データでのファイルサイズを算出する。

$$\text{[全件ファイルサイズ]} (\text{byte}) = \text{[ファイルサイズ単価]} (\text{byte/件}) \times \text{[全件データ]} (\text{件})$$

測定結果から性能単価を算出する。

$$\text{[性能単価]} (\text{秒/件}) = \text{[処理時間]} (\text{秒}) \div 10,000 (\text{件})$$

性能単価から全件データでの蓄積処理時間を算出する。

$$\text{[全件蓄積処理性能]} (\text{秒}) = \text{[性能単価]} (\text{秒/件}) \times \text{[全件データ]} (\text{件})$$

(c) 分散数が異なる場合の WAM 蓄積の処理性能算出

入出力条件が同じ場合で分散数が異なる場合に、性能単価に影響があるか検証する。

検証パターンを表 6-6-3-1 に示す。

表 6-6-3-1. 分散数の違いによる性能測定パターン

#	分散数	分散内訳
1	1	分散サーバ 1 の分散数を 1 とする。
2	2	分散サーバ 1 の分散数を 1 に、分散サーバ 2 の分散数を 1 とする。
3	4	分散サーバ 1 の分散数を 2 に、分散サーバ 2 の分散数を 2 とする。
4	8	分散サーバ 1 の分散数を 4 に、分散サーバ 2 の分散数を 4 とする。

(d) サーバリソース使用状況の取得

GETA の構成が表 6-6-3-1 の #2 の場合の CPU 使用率、I/O wait 発生率、メモリ使用率をサーバリソース使用状況から取得する。

6-6-4. 測定結果と考察

(1) 頻度ファイル、補助ファイル蓄積の性能検証

(a) 測定結果

頻度ファイル、補助ファイル蓄積処理のファイルサイズを表6-6-4-1に、性能単価を表6-6-4-2に示す。

表6-6-4-1. 頻度ファイル、補助ファイルサイズ

#	区分	対象	出力件数	ファイルサイズ	サイズ単価	
1	頻度ファイル	公開公報単語頻度ファイル	全文	10,000 件	45,180,640 byte	4,518 byte
2			要約	10,000 件	4,541,662 byte	454 byte
3			請求項	10,000 件	6,235,183 byte	624 byte
4			請求項 1 *1	9,997 件	3,477,026 byte	348 byte
5			実施例 *1	9,744 件	5,856,846 byte	586 byte
6		公表公報単語頻度ファイル	全文	10,000 件	111,105,073 byte	11,111 byte
7			要約	10,000 件	4,586,830 byte	459 byte
8			請求項	10,000 件	17,181,619 byte	1,718 byte
9			請求項 1 *1	9,823 件	5,142,948 byte	514 byte
10			実施例 *1	6,011 件	40,761,065 byte	4,076 byte
11	頻度ファイル	概念検索頻度ファイル(全文)		10,000 件	50,656,819 byte	5,066 byte
12		概念検索頻度ファイル(要約)		10,000 件	10,017,841 byte	1,002 byte
13		概念検索頻度ファイル(請求項)		10,000 件	11,711,362 byte	1,171 byte
14		概念検索頻度ファイル(請求項 1)		10,000 件	8,953,205 byte	895 byte
15		概念検索頻度ファイル(実施例)		10,000 件	11,333,025 byte	1,133 byte
16		概念検索特定箇所 N 倍頻度ファイル(要約)		10,000 件	50,845,472 byte	5,085 byte
17		概念検索特定箇所 N 倍頻度ファイル(請求項)		10,000 件	50,998,603 byte	5,100 byte
18		概念検索特定箇所 N 倍頻度ファイル(請求項 1)		10,000 件	50,781,991 byte	5,078 byte
19		概念検索特定箇所 N 倍頻度ファイル(実施例)		10,000 件	51,044,939 byte	5,104 byte
20		共起関連分類頻度ファイル		10,000 件	4,771,568 byte	477 byte
21		関連発明者頻度ファイル		10,000 件	1,167,841 byte	117 byte
22		本願別検索式履歴ファイル *2		200 件	262,838 byte	26 byte
23		印刷履歴活用本願別検索式履歴ファイル *2		2 件	16,863 byte	2 byte
23		関連検索キー頻度ファイル		10,000 件	1,927,632 byte	193 byte
24		検索式履歴ワード頻度ファイル		10,000 件	1,038,195 byte	104 byte
25		検索式履歴分類頻度ファイル		10,000 件	1,028,836 byte	103 byte
26		審査関連情報頻度ファイル		10,000 件	50,476,047 byte	5,048 byte
27		補助ファイル	補助ファイル(共起関連分類)		10,000 件	668,777 byte
28	補助ファイル(関連発明者)		10,000 件	668,576 byte	67 byte	
29	補助ファイル(関連検索キー)		10,000 件	681,180 byte	68 byte	
30	補助ファイル(検索履歴ワード)		10,000 件	681,180 byte	68 byte	
31	補助ファイル(検索履歴分類)		10,000 件	681,180 byte	68 byte	
32	補助ファイル(審査関連情報)		10,000 件	710,789 byte	71 byte	
33	補助ファイル(公開公報全文)		10,000 件	710,789 byte	71 byte	
34	補助ファイル(公開公報要約)		10,000 件	710,789 byte	71 byte	

表 6-6-4-1. 頻度ファイル、補助ファイルサイズ（続き）

#	区分	対象	出力件数	ファイルサイズ	サイズ単価
35		補助ファイル(公開公報請求項)	10,000 件	710,789 byte	71 byte
36		補助ファイル(公表公報全文)	10,000 件	945,189 byte	95 byte
37		補助ファイル(公表公報要約)	10,000 件	945,189 byte	95 byte
38		補助ファイル(公表公報請求項)	10,000 件	945,189 byte	95 byte

*1 入力件数に対し出力件数が減少しているが、出願の記載方法が特定のパターンに一致せず抽出できなかったのが原因である。

*2 入力件数に対し出力件数が減少しているが、本願に対応する検索式履歴が存在しなかったのが原因である。

表 6-6-4-2. 頻度ファイル、補助ファイル蓄積処理性能単価

#	区分	対象	出力件数	処理時間	性能単価	
1	頻度ファイル	公開公報単語頻度ファイル	全文	946 秒	0.0946 秒/件	
2			要約			10,000 件
3			請求項			10,000 件
4			請求項 1 *1			9,997 件
5			実施例 *1			9,744 件
6		公表公報単語頻度ファイル	全文	2,369 秒	0.2369 秒/件	
7			要約			10,000 件
8			請求項			10,000 件
9			請求項 1 *1			9,823 件
10			実施例 *1			6,011 件
11		概念検索頻度ファイル(全文)	10,000 件	57 秒	0.0057 秒/件	
12		概念検索頻度ファイル(要約)	10,000 件	33 秒	0.0033 秒/件	
13		概念検索頻度ファイル(請求項)	10,000 件	38 秒	0.0038 秒/件	
14		概念検索頻度ファイル(請求項 1)	10,000 件	36 秒	0.0036 秒/件	
15		概念検索頻度ファイル(実施例)	10,000 件	37 秒	0.0037 秒/件	
16		概念検索特定箇所 N 倍頻度ファイル(要約)	10,000 件	85 秒	0.0085 秒/件	
17		概念検索特定箇所 N 倍頻度ファイル(請求項)	10,000 件	86 秒	0.0086 秒/件	
18		概念検索特定箇所 N 倍頻度ファイル(請求項 1)	10,000 件	88 秒	0.0088 秒/件	
19		概念検索特定箇所 N 倍頻度ファイル(実施例)	10,000 件	86 秒	0.0086 秒/件	
20		共起関連分類頻度ファイル	10,000 件	17 秒	0.0017 秒/件	
21		関連発明者頻度ファイル	10,000 件	37 秒	0.0037 秒/件	
22		本願別検索式履歴ファイル *2	200 件	58 秒	0.0058 秒/件	
23		印刷履歴活用本願別検索式履歴ファイル *2	2 件	24 秒	0.0024 秒/件	
23		関連検索キー頻度ファイル	10,000 件	21 秒	0.0021 秒/件	
24		検索式履歴ワード頻度ファイル	10,000 件	20 秒	0.002 秒/件	
25		検索式履歴分類頻度ファイル	10,000 件	20 秒	0.002 秒/件	
26		審査関連情報頻度ファイル	10,000 件	996 秒	0.0996 秒/件	
27	補助ファイル	補助ファイル(共起関連分類)	10,000 件	19 秒	0.0019 秒/件	
28		補助ファイル(関連発明者)	10,000 件	19 秒	0.0019 秒/件	
29		補助ファイル(関連検索キー)	10,000 件	28 秒	0.0028 秒/件	
30		補助ファイル(検索履歴ワード)	10,000 件	18 秒	0.0018 秒/件	
31		補助ファイル(検索履歴分類)	10,000 件	19 秒	0.0019 秒/件	
32		補助ファイル(審査関連情報)	10,000 件	20 秒	0.002 秒/件	
33		補助ファイル(公開公報全文)	10,000 件	19 秒	0.0019 秒/件	
34		補助ファイル(公開公報要約)	10,000 件	18 秒	0.0018 秒/件	
35		補助ファイル(公開公報請求項)	10,000 件	19 秒	0.0019 秒/件	
36		補助ファイル(公表公報全文)	10,000 件	22 秒	0.0022 秒/件	
37		補助ファイル(公表公報要約)	10,000 件	18 秒	0.0018 秒/件	
38		補助ファイル(公表公報請求項)	10,000 件	18 秒	0.0018 秒/件	

*1 入力件数に対し出力件数が減少しているが、出願の記載方法が特定のパターンに一致せず抽出できなかったのが原因である。

*2 入力件数に対し出力件数が減少しているが、本願に対応する検索式履歴が存在しなかったのが原因である。

(b) 全件データでのファイルサイズ及び蓄積処理性能の算出

表6-6-4-3に示す全件データの件数内訳から、全件データでの頻度ファイル、補助ファイルサイズ、蓄積処理性能を机上検証した結果を表6-6-4-4及び表6-6-4-5に示す。

表6-6-4-3. 全件データの件数内訳

#	全件データ	内訳	年数	範囲	件数
1	公開公報	公開公報	全件	2009年2月時点	6,038,570件
2		公告・登録公報			2,214,732件
3		二次文献			6,931,551件
4		バック分公開公報			2,237,258件
5		公開公報メモ			27,171件
6		登録公報メモ			381件
7		二次文献メモ			445件
8	公表公報	公表公報	全件	2009年2月時点	411,749件
9		バック分公表公報			32,003件
10		PCT-R0 文献			105,375件
11		PCT19 条補正書			788件
12	検索式履歴	検索式履歴	全件	2005年1月から2008年5月 ※本願のある検索式履歴	41,378,305件

表 6-6-4-4. 全件データの頻度ファイル、補助ファイルサイズ

#	区分	対象	データ	サイズ単価	全件件数	全件サイズ	
1		公開公報単語頻度ファイル	全文	公開公報	17,450,108 件	75,188 MB	
2			要約			454 byte	7,558 MB
3			請求項			624 byte	10,376 MB
4			請求項1			348 byte	5,786 MB
5			実施例			586 byte	9,747 MB
6		公表公報単語頻度ファイル	全文	公表公報	549,915 件	5,827 MB	
7			要約			459 byte	241 MB
8			請求項			1,718 byte	901 MB
9			請求項1			514 byte	270 MB
10			実施例			4,076 byte	2,138 MB
11	頻度ファイル	概念検索頻度ファイル(全文)		公開公報 公表公報	18,000,023 件	86,958 MB	
12		概念検索頻度ファイル(要約)		公開公報 公表公報	18,000,023 件	17,197 MB	
13		概念検索頻度ファイル(請求項)		公開公報 公表公報	18,000,023 件	20,104 MB	
14		概念検索頻度ファイル(請求項1)		公開公報 公表公報	18,000,023 件	15,369 MB	
15		概念検索頻度ファイル(実施例)		公開公報 公表公報	18,000,023 件	19,454 MB	
16		概念検索特定箇所N倍 頻度ファイル(要約)		公開公報 公表公報	18,000,023 件	87,282 MB	
17		概念検索特定箇所N倍 頻度ファイル(請求項)		公開公報 公表公報	18,000,023 件	87,545 MB	
18		概念検索特定箇所N倍 頻度ファイル(請求項1)		公開公報 公表公報	18,000,023 件	87,173 MB	
19		概念検索特定箇所N倍 頻度ファイル(実施例)		公開公報 公表公報	18,000,023 件	87,625 MB	
20		共起関連分類頻度ファイル		公開公報 公表公報	18,000,023 件	8,191 MB	
21	関連発明者頻度ファイル		公開公報 公表公報	18,000,023 件	2,005 MB		
22	本願別検索式履歴ファイル		検索式履歴	26 byte	41,378,305 件	1,037 MB	
23	印刷履歴活用本願別検索式履歴ファイル		検索式履歴	2 byte	41,378,305 件	67 MB	
24	関連検索キー頻度ファイル		検索式履歴	193 byte	41,378,305 件	7,607 MB	
25	検索式履歴ワード頻度ファイル		検索式履歴	104 byte	41,378,305 件	4,097 MB	
26	検索式履歴分類頻度ファイル		検索式履歴	103 byte	41,378,305 件	4,060 MB	
27	審査関連情報頻度ファイル		公開公報 公表公報	5,048 byte	18,000,023 件	86,648 MB	
28	補助ファイル	補助ファイル(共起関連分類)		公開公報 公表公報	18,000,023 件	1,148 MB	
29		補助ファイル(関連発明者)		公開公報 公表公報	18,000,023 件	1,148 MB	

表 6-6-4-4. 全件データの頻度ファイル、補助ファイルサイズ（続き）

#	区分	対象	データ	サイズ単価	全件件数	全件サイズ
30		補助ファイル(関連検索キー)	検索式履歴	68 byte	41,378,305 件	2,688 MB
31		補助ファイル(検索履歴ワード)	検索式履歴	68 byte	41,378,305 件	2,688 MB
32		補助ファイル(検索履歴分類)	検索式履歴	68 byte	41,378,305 件	2,688 MB
33		補助ファイル(審査関連情報)	公開公報 公表公報	71 byte	18,000,023 件	1,220 MB
34		補助ファイル(公開公報全文)	公開公報	71 byte	17,450,108 件	1,183 MB
35		補助ファイル(公開公報要約)	公開公報	71 byte	17,450,108 件	1,183 MB
36		補助ファイル(公開公報請求項)	公開公報	71 byte	17,450,108 件	1,183 MB
37		補助ファイル(公表公報全文)	公表公報	95 byte	549,915 件	50 MB
38		補助ファイル(公表公報要約)	公表公報	95 byte	549,915 件	50 MB
39		補助ファイル(公表公報請求項)	公表公報	95 byte	549,915 件	50 MB
合計					673,026,877 件	755,730 MB

表6-6-4-5. 全件データの頻度ファイル、補助ファイル蓄積処理性能

#	区分	対象	データ	性能単価	全件件数	全件性能	
1	頻度ファイル	公開公報単語頻度ファイル	公開公報	0.0946 秒/件	17,450,108 件	459 時間	
2							全文
3							要約
4							請求項
5							請求項1
6		公表公報単語頻度ファイル	公表公報	0.2369 秒/件	549,915 件	37 時間	
7							全文
8							要約
9							請求項
10							請求項1
11		概念検索頻度ファイル(全文)		公開公報 公表公報	0.0057 秒/件	18,000,023 件	29 時間
12		概念検索頻度ファイル(要約)		公開公報 公表公報	0.0033 秒/件	18,000,023 件	17 時間
13		概念検索頻度ファイル(請求項)		公開公報 公表公報	0.0038 秒/件	18,000,023 件	20 時間
14		概念検索頻度ファイル(請求項1)		公開公報 公表公報	0.0036 秒/件	18,000,023 件	19 時間
15		概念検索頻度ファイル(実施例)		公開公報 公表公報	0.0037 秒/件	18,000,023 件	19 時間
16		概念検索特定箇所N倍 頻度ファイル(要約)		公開公報 公表公報	0.0085 秒/件	18,000,023 件	43 時間
17		概念検索特定箇所N倍 頻度ファイル(請求項)		公開公報 公表公報	0.0086 秒/件	18,000,023 件	44 時間
18		概念検索特定箇所N倍 頻度ファイル(請求項1)		公開公報 公表公報	0.0088 秒/件	18,000,023 件	45 時間
19		概念検索特定箇所N倍 頻度ファイル(実施例)		公開公報 公表公報	0.0086 秒/件	18,000,023 件	44 時間
20		共起関連分類頻度ファイル		公開公報 公表公報	0.0017 秒/件	18,000,023 件	9 時間
21		関連発明者頻度ファイル		公開公報 公表公報	0.0037 秒/件	18,000,023 件	19 時間
22		本願別検索式履歴ファイル		検索式履歴	0.0058 秒/件	41,378,305 件	67 時間
23		印刷履歴活用本願別検索式履歴ファイル		検索式履歴	0.0024 秒/件	41,378,305 件	28 時間
24		関連検索キー頻度ファイル		検索式履歴	0.0021 秒/件	41,378,305 件	25 時間
25		検索式履歴ワード頻度ファイル		検索式履歴	0.002 秒/件	41,378,305 件	23 時間
26		検索式履歴分類頻度ファイル		検索式履歴	0.002 秒/件	41,378,305 件	23 時間
27		審査関連情報頻度ファイル		公開公報 公表公報	0.0996 秒/件	18,000,023 件	499 時間
28	補助ファイル	補助ファイル(共起関連分類)	公開公報 公表公報	0.0019 秒/件	18,000,023 件	10 時間	
29		補助ファイル(関連発明者)	公開公報 公表公報	0.0019 秒/件	18,000,023 件	10 時間	

表 6-6-4-5. 全件データの頻度ファイル、補助ファイル蓄積処理性能（続き）

#	区分	対象	データ	性能単価	全件件数	全件性能
30	補助 フ ァ ィ ル	補助ファイル(関連検索キー)	検索式履歴	0.0028 秒/件	41,378,305 件	33 時間
31		補助ファイル(検索履歴ワード)	検索式履歴	0.0018 秒/件	41,378,305 件	21 時間
32		補助ファイル(検索履歴分類)	検索式履歴	0.0019 秒/件	41,378,305 件	22 時間
33		補助ファイル(審査関連情報)	公開公報 公表公報	0.002 秒/件	18,000,023 件	11 時間
34		補助ファイル(公開公報全文)	公開公報	0.0019 秒/件	17,450,108 件	10 時間
35		補助ファイル(公開公報要約)	公開公報	0.0018 秒/件	17,450,108 件	9 時間
36		補助ファイル(公開公報請求項)	公開公報	0.0019 秒/件	17,450,108 件	10 時間
37		補助ファイル(公表公報全文)	公表公報	0.0022 秒/件	549,915 件	1 時間
38		補助ファイル(公表公報要約)	公表公報	0.0018 秒/件	549,915 件	1 時間
39		補助ファイル(公表公報請求項)	公表公報	0.0018 秒/件	549,915 件	1 時間
合計					673,026,877 件	1608 時間

(c) 頻度ファイル、補助ファイルの蓄積処理時のサーバリソース

頻度ファイル、補助ファイルの蓄積処理時の CPU 使用率と I/O wait 発生率を図 6-6-4-1 に、メモリ使用量を図 6-6-4-2 に示す。

図 6-6-4-1 から以下のことが言える。

- ・ CPU 使用率(1 コア)は常に 100%である。
- ・ I/O wait は発生しない。

図 6-6-4-2 から以下のことが言える。

- ・ メモリの平均使用率は 15%であり、急激なメモリの増減は発生していない。

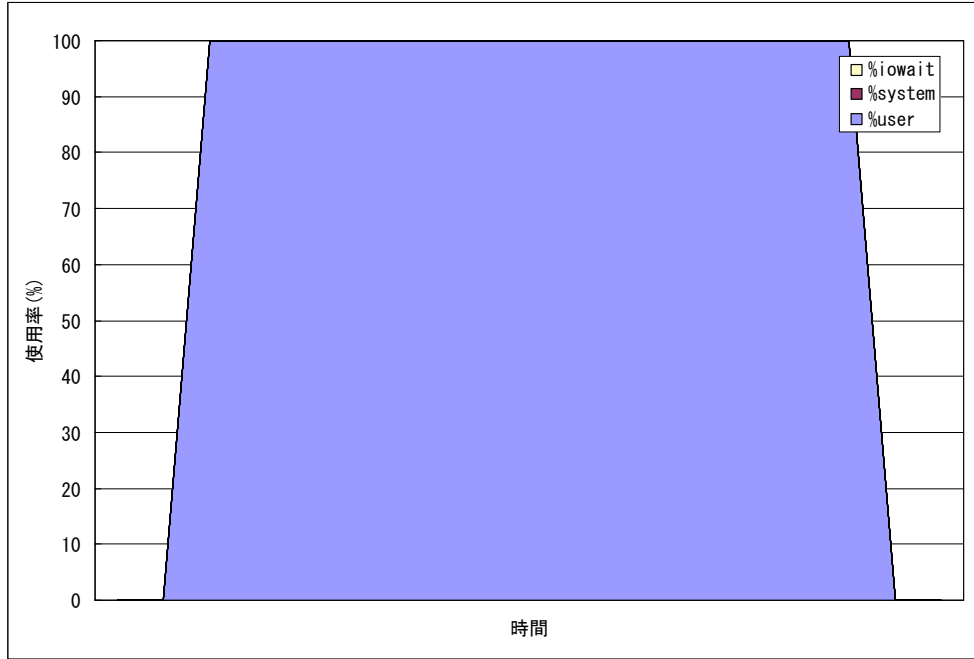


図 6-6-4-1. CPU 使用率、I/O wait 発生率

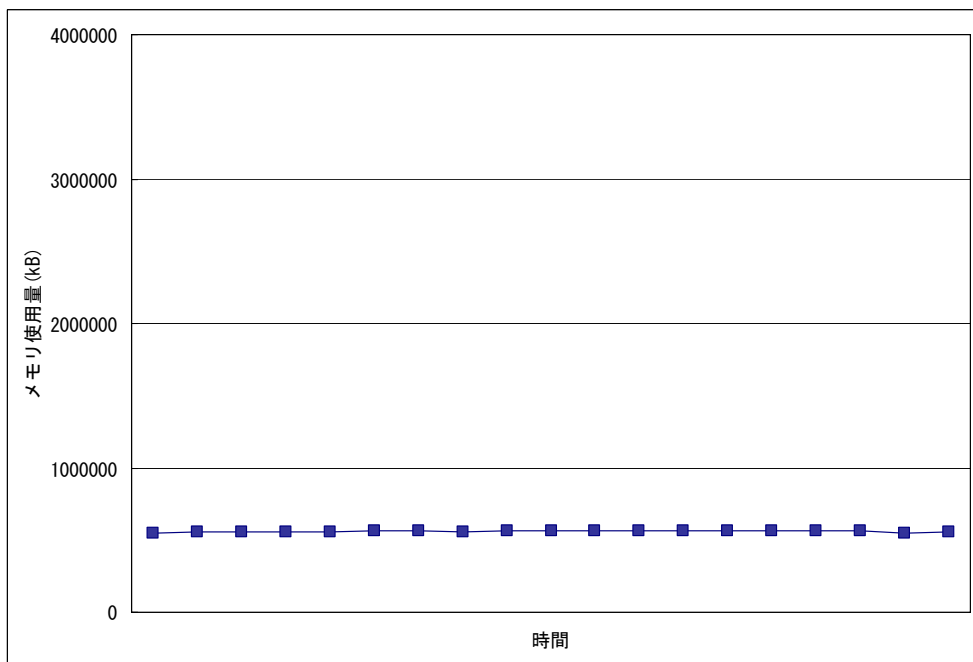


図 6-6-4-2. メモリ使用量

(d) 考察

全件データでの頻度ファイル、補助ファイル蓄積処理性能について検証した結果を報告する。

(i) 単語頻度ファイルの蓄積処理性能

公開公報単語頻度ファイル及び公表公報単語頻度ファイルの蓄積処理は、公報の文章からワードを切り出す形態素解析を行い、同結果からワードと頻度を算出するフルテキストに対する処理であるため他ファイルに比べて比較的処理時間がかかる。

頻度ファイル、補助ファイルは公報や検索履歴単位に作成するため、多重で作成することが可能である。そのためサーバ台数の増加により蓄積時間の短縮が可能である。

(ii) 蓄積時のサーバリソース

- ・ CPU 利用率(1 コア)は常に 100%である。
- ・ I/O wait は発生しない。
- ・ メモリの平均使用率は 15%であり、急激なメモリの増減は発生していない。

以上のことからマシンの能力は十分であったといえる。

(2) WAM 蓄積

(a) 測定結果

WAM 蓄積処理のファイルサイズを表 6-6-4-6 に、性能単価を表 6-6-4-7 に示す。

表 6-6-4-6. WAM サイズ

#	入力ファイル		出力件数	ファイルサイズ	サイズ単価
1	公開公報	全文	10,000 件	23,338,948 byte	2,334 byte
2		要約	10,000 件	2,585,091 byte	259 byte
3		請求項	10,000 件	3,559,533 byte	356 byte
4	公表公報	全文	10,000 件	55,583,211 byte	5,558 byte
5		要約	10,000 件	2,589,934 byte	259 byte
6		請求項	10,000 件	9,181,695 byte	918 byte
7	概念検索	全文	10,000 件	27,552,235 byte	2,755 byte
8		要約	10,000 件	6,676,092 byte	668 byte
9		請求項	10,000 件	7,653,882 byte	765 byte
10		請求項 1	10,000 件	6,137,044 byte	614 byte
11		実施例	10,000 件	7,483,618 byte	748 byte
12		要約 N 倍	10,000 件	28,045,108 byte	2,805 byte
13		請求項 N 倍	10,000 件	28,725,031 byte	2,873 byte
14		請求項 1N 倍	10,000 件	27,941,315 byte	2,794 byte
15		実施例 N 倍	10,000 件	28,484,690 byte	2,848 byte
16	データマイニング	共起関連分類	10,000 件	4,181,412 byte	418 byte
17		関連発明者	10,000 件	927,322 byte	93 byte
18		審査関連情報	10,000 件	2,595,633 byte	2,785 byte
19		関連検索キー	10,000 件	1,216,976 byte	260 byte
20		検索式履歴ワード	10,000 件	1,616,914 byte	122 byte
21		検索式履歴分類	10,000 件	27,845,829 byte	162 byte

表 6-6-4-7. WAM 蓄積処理性能単価

#	入力ファイル		出力要素数	出力件数	処理時間	性能単価
1	公開公報	全文	138,123 語	10,000 件	9 秒	0.0009 秒
2		要約	18,671 語	10,000 件	1 秒	0.0001 秒
3		請求項	21,068 語	10,000 件	1 秒	0.0001 秒
4	公表公報	全文	501,247 語	10,000 件	26 秒	0.0026 秒
5		要約	22,200 語	10,000 件	1 秒	0.0001 秒
6		請求項	73,154 語	10,000 件	3 秒	0.0003 秒
7	概念検索	全文	244,101 語	10,000 件	11 秒	0.0011 秒
8		要約	124,649 語	10,000 件	3 秒	0.0003 秒
9		請求項	127,046 語	10,000 件	3 秒	0.0003 秒
10		請求項 1	120,804 語	10,000 件	2 秒	0.0002 秒
11		実施例	133,689 語	10,000 件	3 秒	0.0003 秒
12		要約	244,101 語	10,000 件	11 秒	0.0011 秒
13		請求項	244,101 語	10,000 件	11 秒	0.0011 秒
14		請求項 1	244,101 語	10,000 件	11 秒	0.0011 秒
15		実施例	244,101 語	10,000 件	11 秒	0.0011 秒
16	データマイニング	共起関連分類	107,915 語	10,000 件	2 秒	0.0002 秒
17		関連発明者	20,242 語	10,000 件	1 秒	0.0001 秒
18		審査関連情報	258,371 語	10,000 件	11 秒	0.0011 秒
19		関連検索キー	67,870 語	10,000 件	2 秒	0.0002 秒
20		検索式履歴ワード	26,789 語	10,000 件	5 秒	0.0005 秒
21		検索式履歴分類	41,054 語	10,000 件	2 秒	0.0002 秒

(b) 全件データでのファイルサイズ及び蓄積処理性能の算出

表6-6-4-8に示す全件データの件数内訳から、全件データでのWAMサイズ、蓄積処理性能を机上検証した結果を表6-6-4-9及び表6-6-4-10に示す。

表6-6-4-8. 全文テキストデータ全件の件数内訳

#	全件データ	内訳	年数	範囲	件数
1	公開公報	公開公報	全件	2009年2月時点	6,038,570件
2		公告・登録公報			2,214,732件
3		二次文献			6,931,551件
4		バック分公開公報			2,237,258件
5		公開公報メモ			27,171件
6		登録公報メモ			381件
7		二次文献メモ			445件
8	公表公報	公表公報	全件	2009年2月時点	411,749件
9		バック分公表公報			32,003件
10		PCT-R0 文献			105,375件
11		PCT19 条補正書			788件
12	検索式履歴	検索式履歴	全件	2005年1月から2008年5月 ※本願のある検索式履歴	41,378,305件

表6-6-4-9. 全件データのWAM頻度ファイル、補助ファイルサイズ

#	入力ファイル	データ	サイズ単価	全件件数	全件サイズ
1	公開公報	全文	2,334 byte	17,450,108件	38,841 MB
2		要約	259 byte	17,450,108件	4,303 MB
3		請求項	356 byte	17,450,108件	5,924 MB
4	公表公報	全文	5,558 byte	549,915件	2,916 MB
5		要約	259 byte	549,915件	136 MB
6		請求項	918 byte	549,915件	482 MB
7	概念検索	全文	2,755 byte	18,000,023件	47,297 MB
8		要約	668 byte	18,000,023件	11,461 MB
9		請求項	765 byte	18,000,023件	13,139 MB
10		請求項1	614 byte	18,000,023件	10,535 MB
11		実施例	748 byte	18,000,023件	12,847 MB
12		要約N倍	2,805 byte	18,000,023件	48,143 MB
13		請求項N倍	2,873 byte	18,000,023件	49,310 MB
14		請求項1N倍	2,794 byte	18,000,023件	47,965 MB
15		実施例N倍	2,848 byte	18,000,023件	48,898 MB
16	データマイニング	共起関連分類	418 byte	18,000,023件	7,178 MB
17		関連発明者	93 byte	18,000,023件	1,592 MB
18		審査関連情報	2,785 byte	18,000,023件	10,243 MB
19		関連検索キー	260 byte	41,378,305件	4,803 MB
20		検索式履歴ワード	122 byte	41,378,305件	6,381 MB
21		検索式履歴分類	162 byte	41,378,305件	47,801 MB
合計				394,135,260件	420,195 MB

表 6-6-4-10. 全件データの WAM 蓄積処理性能

#	入力ファイル		データ	性能単価	全件件数	全件性能
1	公開公報	全文	公開公報	0.0009 秒	17,450,108 件	5 時間
2		要約		0.0001 秒	17,450,108 件	1 時間
3		請求項		0.0001 秒	17,450,108 件	1 時間
4	公表公報	全文	公表公報	0.0026 秒	549,915 件	1 時間
5		要約		0.0001 秒	549,915 件	1 時間
6		請求項		0.0003 秒	549,915 件	1 時間
7	概念検索	全文	公開公報	0.0011 秒	18,000,023 件	6 時間
8		要約	公表公報	0.0003 秒	18,000,023 件	2 時間
9		請求項		0.0003 秒	18,000,023 件	2 時間
10		請求項 1		0.0002 秒	18,000,023 件	2 時間
11		実施例		0.0003 秒	18,000,023 件	2 時間
12		要約		0.0011 秒	18,000,023 件	6 時間
13		請求項		0.0011 秒	18,000,023 件	6 時間
14		請求項 1		0.0011 秒	18,000,023 件	6 時間
15		実施例		0.0011 秒	18,000,023 件	6 時間
16	データマイニング	共起関連分類	公開公報	0.0002 秒	18,000,023 件	2 時間
17		関連発明者	公表公報	0.0001 秒	18,000,023 件	1 時間
18		審査関連情報		0.0011 秒	18,000,023 件	3 時間
19		関連検索キー	検索式履歴	0.0002 秒	41,378,305 件	6 時間
20		検索式履歴ワード		0.0005 秒	41,378,305 件	3 時間
21		検索式履歴分類		0.0002 秒	41,378,305 件	6 時間
合計					394,135,260 件	69 時間

(c) 分散数が異なる場合の WAM 蓄積時間の検証

入出力条件が同じ場合で分散数が異なる場合、WAM の作成時間に影響があるか検証を行う。

入出力条件が表 6-6-4-1 1 の場合に、WAM の蓄積性能の測定を行い、算出した性能単価を表 6-6-4-1 2 および図 6-6-4-3 に示す。

検証の結果、分散数が増加しても、WAM の蓄積処理時間はほとんど変動しないことが分かった。

表 6-6-4-1 1. 入出力条件

#	入力ファイル		ファイルサイズ	入力件数	出力件数	出力単語数
1	公表公報	全文	頻度ファイル	111,105,073 byte	10,000 件	501,247 語
2			タイトルファイル	945,189 byte		

表 6-6-4-1 2. 分散数ごとの性能単価

#	分散数	入力件数	1分散あたりの件数	WAM 作成時間	性能単価
1	1	10,000 件	10,000 件	25 秒	0.0025 秒
2	2	10,000 件	5,000 件	26 秒	0.0026 秒
3	4	10,000 件	2,500 件	27 秒	0.0027 秒
4	8	10,000 件	1,250 件	28 秒	0.0028 秒

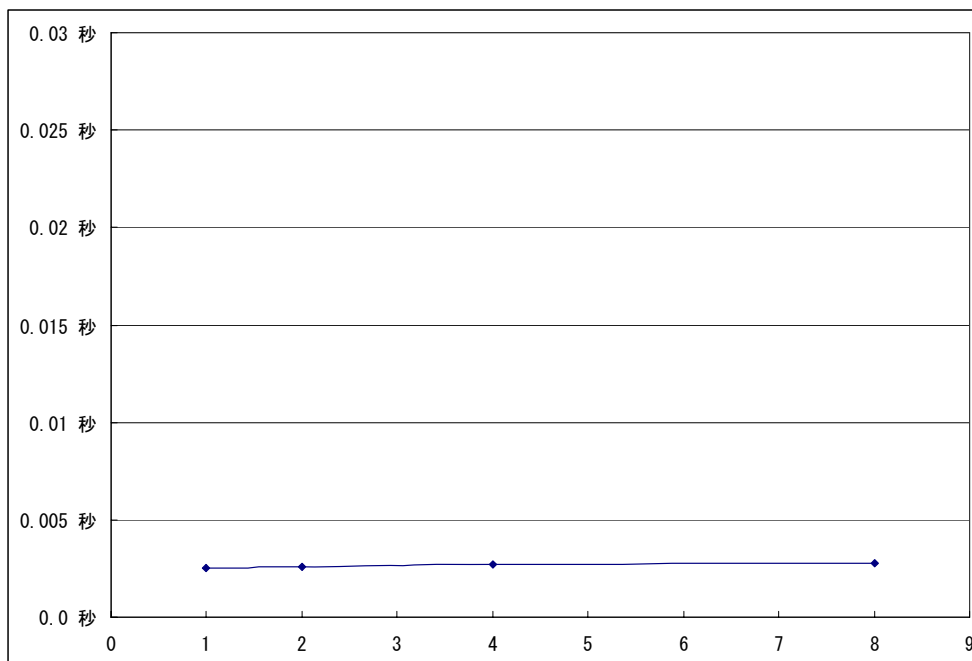


図 6-6-4-3. 分散数による性能単価の推移

(d) WAM の蓄積処理時のサーバリソース

頻度ファイル、補助ファイルの蓄積処理時の CPU 使用率と I/O wait 発生率を図 6-6-4-4、図 6-6-4-5、図 6-6-4-6、図 6-6-4-7 に、メモリ使用量を図 6-6-4-8 に示す。

図 6-6-4-4、図 6-6-4-5、図 6-6-4-6、図 6-6-4-7 から以下のことが言える。

- ・ AP サーバの CPU 使用率(2 コア)は常に 100%で、DB サーバの CPU 使用率は最大で 50%である。
- ・ I/O wait の発生率は極めて低い。

図 6-6-4-8 から以下のことが言える。

- ・ メモリの平均使用率は AP サーバが 6%、DB サーバが共に 3%であり、急激なメモリの増減は発生していない。

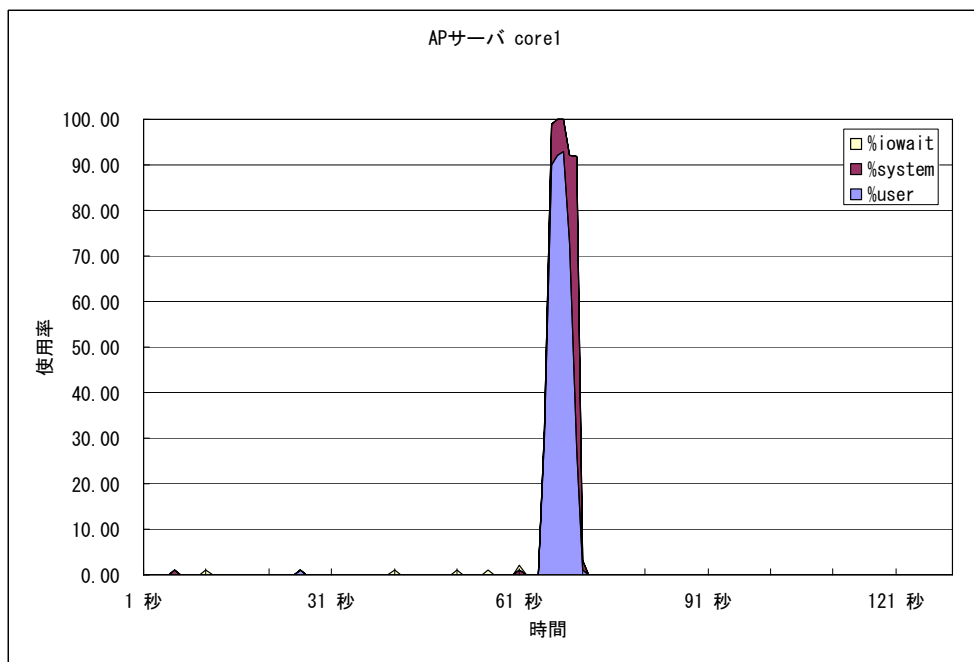


図 6-6-4-4. AP サーバの CPU コア 1 の CPU 使用率、I/O wait 発生率

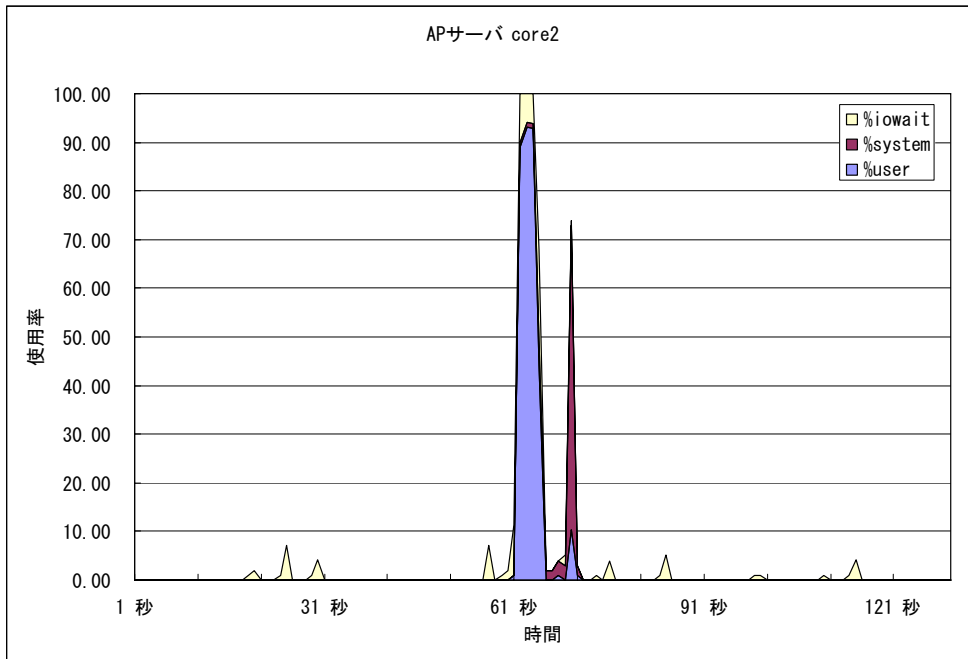


図 6-6-4-5. AP サーバの CPU コア 2 の CPU 使用率、I/O wait 発生率

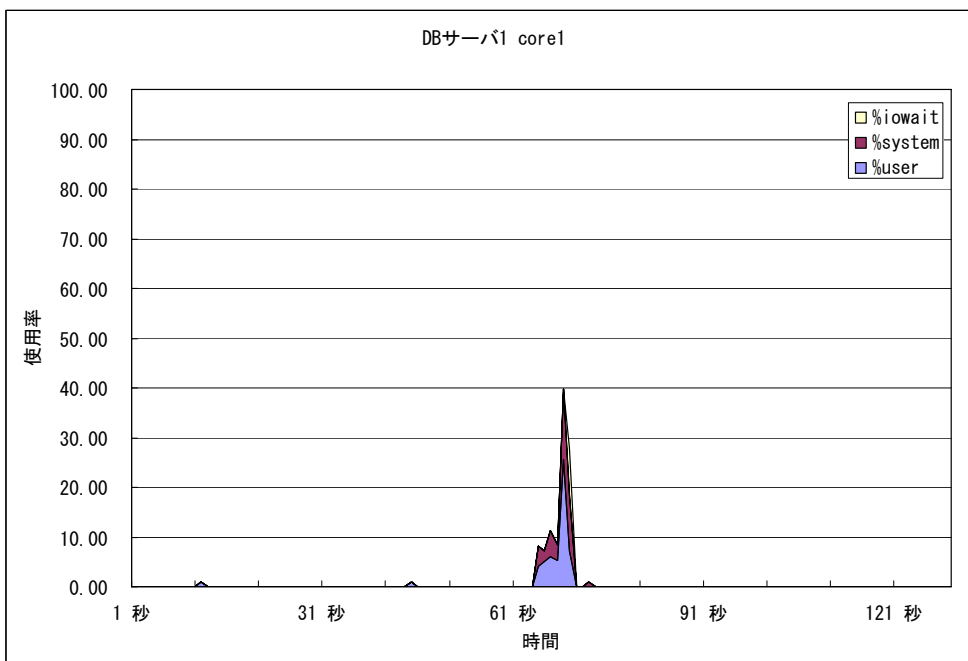


図 6-6-4-6. DB サーバ 1 の CPU コア 1 の CPU 使用率、I/O wait 発生率

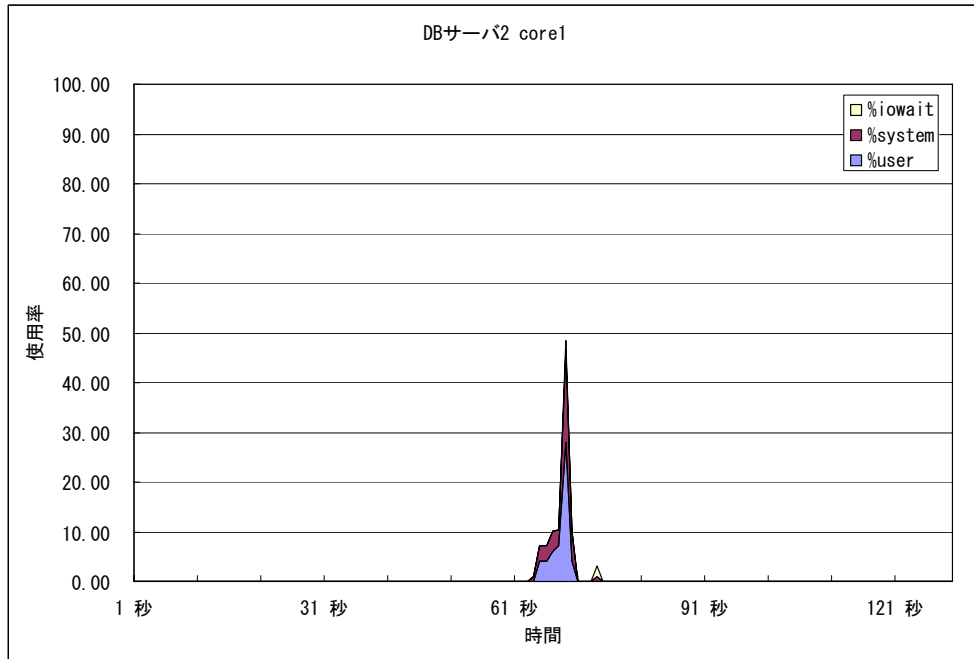


図 6-6-4-7. DBサーバ2のCPUコア1のCPU使用率、I/O wait 発生率

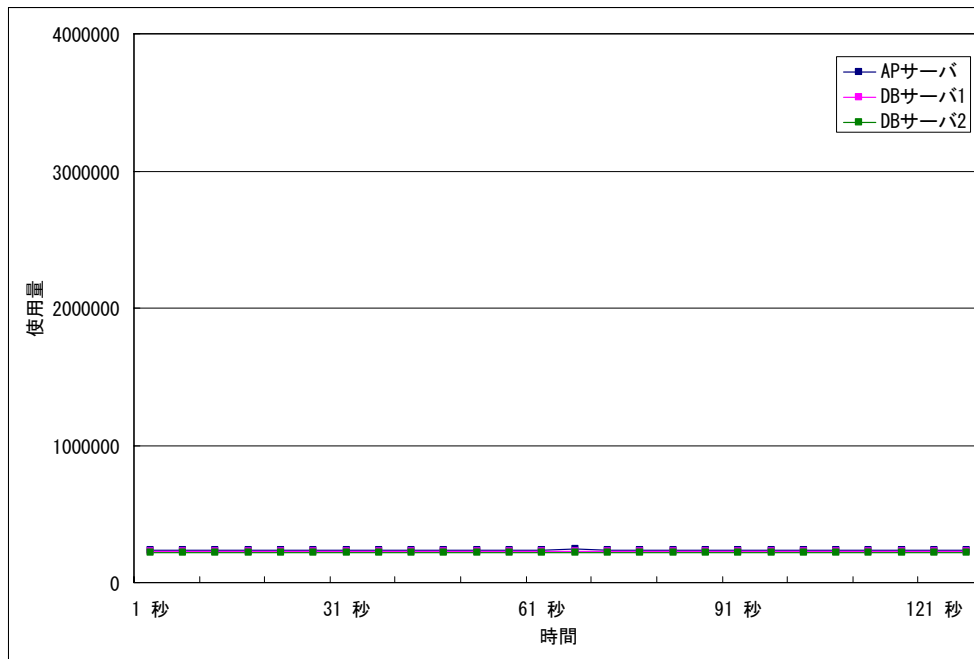


図 6-6-4-8. 各マシンのメモリ使用量

(e) 考察

WAM 蓄積処理について、全件データでの性能を検証した結果を報告する。

(i) WAM の蓄積時間

WAM の作成時間は頻度ファイル、補助ファイルの作成より短時間で完了する。

WAM の分散数による蓄積処理時間への影響はほとんどない。

(ii) サーバリソース

- ・ AP サーバの CPU 使用率(2 コア)は常に 100%で、DB サーバの CPU 使用率は最大で 50%である。
- ・ I/O wait の発生率は極めて低い。
- ・ メモリの平均使用率は AP サーバが 6%、DB サーバが共に 3%であり、急激なメモリの増減は発生していない。

以上のことからマシンの能力は十分であったといえる。

6-6-5. 蓄積のまとめ

データ蓄積についてのまとめを表 6-6-5-1 に示す。

表 6-6-5-1. データ蓄積のまとめ

#	蓄積のまとめ	参照先
1	頻度ファイル、補助ファイルは公報や検索履歴単位に作成するため、多重で作成することが可能である。そのためサーバ台数の増加により蓄積時間の短縮が可能である。	6-6-2 章
2	蓄積処理においてサーバリソースの CPU 利用率が常に 100%であるため、CPU 能力の向上により処理時間の短縮が可能である。	図 6-6-4-1 図 6-6-4-4
3	データを更新する場合、頻度ファイル、補助ファイルは更新分のデータの頻度ファイル、補助ファイルのみを作ればよいため、更新の際の蓄積時間は初回蓄積時より軽減される。 WAMについては最新の GETA バージョンではデータの追加更新ができないため既存のデータを含め再作成する必要があるが、全件の蓄積時間が最大 6 時間であり、頻繁に更新がない限り問題はないと言える。	6-6-2 章

6-7. 性能検証のまとめ

(1) オンライン処理

概念検索については、1600万件のテキストデータに対して、一般的なスペックのPCでも実用的なスピードが確保できることが分かった。

データマイニングについては、検証ツールでは目標性能を達成できないが、スペクトル表示の処理を改善することで、概念検索と同程度まで処理時間の短縮が可能である。

また、テキストデータ1600万件時の必要なサーバ台数を机上検証で求めた結果について、表6-7-1に示す。

表6-7-1. テキストデータ1600万件時の必要なサーバ台数

#	機能	サーバ	庁内環境	庁外環境
1	概念検索	APサーバ	4台	2台
2		DBサーバ	544台	138台
3	データマイニング(参考)	APサーバ	3台	2台
4		DBサーバ	132台	60台
合計			683台	202台

(2) 蓄積処理

データ蓄積ツールの処理時間を測定することで、頻度ファイル、補助ファイル、WAMの蓄積性能単価を算出することができた。また、性能単価から特許庁保有データ全件に対するデータ蓄積処理時間の予測を行うことができた。

参考として、公開公報(全文)のデータ蓄積におけるファイルサイズと処理時間を表6-7-2に示す。

表6-7-2. テキストデータ1600万件時の必要なサーバ台数

#	データ	サーバ	件数	ファイルサイズ	処理時間	並列処理
1	公開公報 (全文)	頻度ファイル	17,450,108件	75,188MB	459時間	可能
2		補助ファイル	17,450,108件	1,183MB	10時間	可能
3		WAM	17,450,108件	38,841MB	5時間	—