

XML 型データベースエンジン「Shunsaku^{注1)}」による特許・実用新案全文検索システム

Full Text Search System for Patents and Utility Models Using the XML-based Database Engine "Shunsaku"

小柳恵介 *
Keisuke Oyanagi

鈴木義昭 *
Yoshiaki Suzuki

田中秀樹 *
Hideki Tanaka

宮地優逸 *
Yuichi Miyachi

渡瀬浩章 *
Hiroaki Watase

松井洋介 *
Yosuke Matsui

野水俊明 *
Toshiaki Nomizu

菅原浩一 *
Kouichi Sugawara

* ソリューションビジネス本部 システム事業部 第一システム統括部 第一システム部

日本特許庁では、特許審査の迅速化を目的とした審査官の大幅増員に伴う「特許・実用新案全文検索システム」のリプレースが計画され、富士通が XML 型データベースエンジン「Shunsaku」を提案・入札し受注した。PFU は、本システムにおいて特許情報の XML 化及び、全文検索を実行する業務アプリケーションの開発を担当した。特許情報を XML 化し全文検索するにあたり、XML データの正規化、データ蓄積方法の最適化といった技術を活用することにより、厳しい顧客要件を満たすことに成功している。

Due to the rapid increase in recruitment of examiners which is aimed at speeding up patent examination procedures, the Japan Patent Office is planning to replace the "full text search system for patents and utility models". As a solution to this plan, Fujitsu Limited has suggested the installation of the XML-based database engine "Shunsaku", and received an order for this system through tender. In installing this system, PFU was responsible for converting patent information into XML data, and developing a business application which executes the full text search operation. In doing so, PFU was successful in meeting the strict requirements of the client by applying technological skills such as normalization of XML data and optimization of data accumulation methods.

1 まえがき

我が国では、2002 年国家戦略として知的財産立国を目指し知的財産戦略大綱が取りまとめられた。そして、知的財産権の保護を強化する「プロパテント政策」が打ち出されて以来、知的財産の創造・保護・活用が重要な課題となっている（図 - 1 参照）。知的財産の創造・保護・活用といった知的創造サイクルの中心にある日本特許庁では、知的財産の迅速で、しかも的確な審査といったサービスの充実が急務となっている。

そして今回、特許審査の迅速化を目的とした審査官の大幅増員に伴う「特許・実用新案全文検索システム（以降、特実全文検索システムと略す）」のリプレースが計画され、調達仕様書が公開された。

本システムのリプレースにあたり、富士通は、XML

型データベースエンジン「Interstage^{注2)}Shunsaku¹⁾ Data Manager（以降、Shunsaku と略す）」を提案

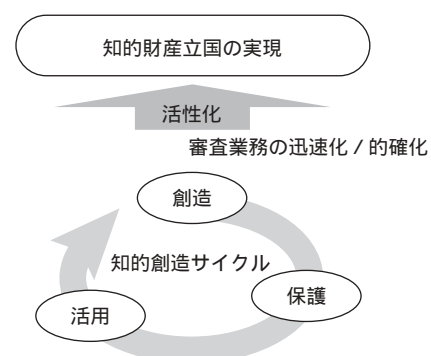


図 1 知的創造サイクルの活性化
(Fig.1-Invigoration of intellectual creativity cycle)

注 1) Shunsaku は富士通株式会社の登録商標である。

注 2) INTERSTAGE は富士通株式会社の登録商標である。

して受注するに至った。

PFU は、本システムの提案・入札作業から参画しており、システム構築作業においても、全文検索を実行する検索アプリケーションと、新規特許情報を XML に変換して Shunsaku に蓄積する登録アプリケーションの開発及び、旧システムからのデータ移行を担当した。

Shunsaku は、「インデックスレスの高速検索」、「並列検索機構」、「シンプルなスケールアップ運用」といった優れた特性を持ったデータベースエンジンである。その反面、その特性を十分に引き出すためには、他のデータベースエンジンでは見られない技術・ノウハウが必要となる。PFU は、これまでに培った Shunsaku 及び XML の技術を活用し、厳しい顧客要件の実現に成功した。

2 特実全文検索システムの概要

2.1 システム概要

特実全文検索システムとは、審査官の特許審査業務において、出願審査請求がなされた出願情報の先行技術調査（類似特許検索）に用いられるシステムである。図 - 2 にシステム概要を示す。

出願者より登録、出願された出願情報及び、特許認定された特許情報（以降、纏めて特許情報と略す）を、検索用データとして XML に変換・Shunsaku 内に格納し、審査官が入力した検索条件を用いて、類似特許の全文検索を実行する。その検索結果として、ヒットした特許情報のキー情報^{注3)}を返す。約 4 800 万件（約 1 TB）の特許情報を検索対象とし、検索結果として最大で 10 万件のキー情報を返す必要がある。

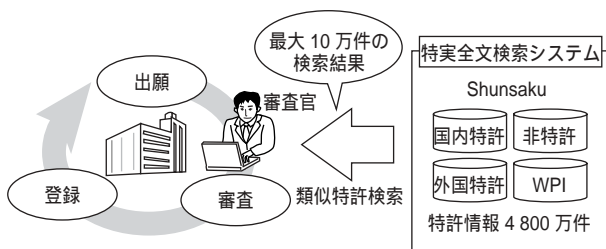


図 2 特実全文検索システムの概要

(Fig.2-Outline of full text search system for patents and utility models)

注 3) 一つの特許情報を識別するためのユニークキー。本システムが返したキー情報を元に、他システム上の表示用データが取得され審査官端末に画面表示される。

2.2 システムの構成

本システムは、複数の AP（アプリケーション）サーバと複数の Shunsaku システムで構成される。また、Shunsaku システムは、特許文献データ単位（国内文献、外国文献他）に存在し、全文検索用のサーチデータは、Shunsaku システムを構成する複数のサーチサーバ（ブレードシステムサーバ^{注4)}）のメモリに分散配置され、全てのサーバを用いた並列検索が同時実行される。コンダクタサーバはディレクタサーバを、ディレクタサーバはサーチサーバをそれぞれ管理するサーバである。

システムの規模としては、1 つの Shunsaku システムあたり、約 20 台～ 200 台のサーバ群で構成され、全体でサーバ数約 500 台規模の大規模なシステムとなっている。図 - 3 にシステムの構成を示す。

2.3 厳しい性能要件

本システムでは、以下の厳しい性能要件を満たす必要があった。

(1) オンライン性能要件

審査官からの検索要求を受け、全文検索を実行するオンライン業務処理の性能要件を表 - 1 に示す。調達仕様書では、業務ピーク時間帯にこの性能要件を満たすことを要求された。

(2) バッチ性能要件

本システムでは、日次で約 130 万件の特許情報の追加・更新を行う。また、Shunsaku へのデータ登録は

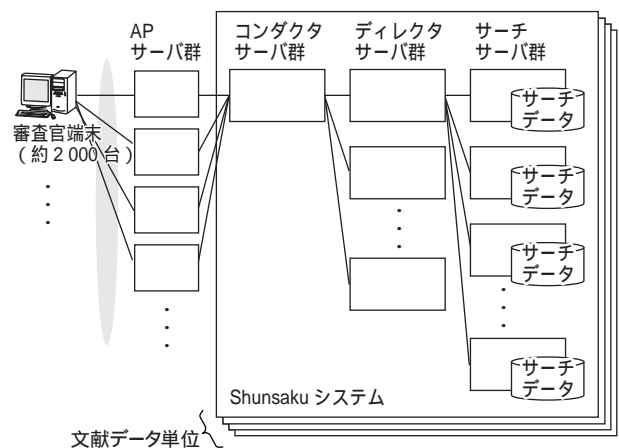


図 3 特実全文検索システムの構成

(Fig.3-Configuration of full text search system for patents and utility models)

注 4) 1 枚の基板にコンピュータとして必要な要素を実装し、必要な枚数を接続して構成するサーバ専用機。本システムでは、1 サーチサーバあたり、7 枚のブレードサーバで構成されている。

オンライン業務時間外となるため、システムの運用スケジュール上、追加・更新処理に許された時間は、2 時間に限られていた。

3 システムの構築にあたって

3.1 課題と開発のポイント

(1) 検索結果マージ処理に要する検索性能への影響

特許情報は、一つの出願につき、そのライフサイクル単位で、同一のキー情報をもつ複数の文献データ（公開公報，登録公報，文献メモ等）で構成される。この文献データそれぞれを、別個に Shunsaku システムに格納した場合、同一のキー情報のマージが必要になる。図 - 4 は特許情報検索結果のマージ処理の様子を示す。

このような処理方式をとった場合、マージ処理によ

表 - 1 オンライン性能要件

ヒット件数 文献種別	10 万件 以下	10 万件	10 万件 以上 ^{注1)}	想定 データ量 (GB)
国内文献	5.0 秒	60.0 秒	5.0 秒	401
外国文献	10.0 秒	-	-	278
非特許文献	5.0 秒	-	-	64
WPI ^{注2)} 文献	5.0 秒	-	-	343
合 計				1 086

注 1) ヒット件数が 10 万件以上の場合、アプリケーションは処理を中止し、エラー応答することが機能要件となっているため、性能要件としては 10 万件以下と同等となっている。

注 2) イギリスのダーウェント社が作成する特許データベースに格納された、31 カ国 2 機関の特許情報。

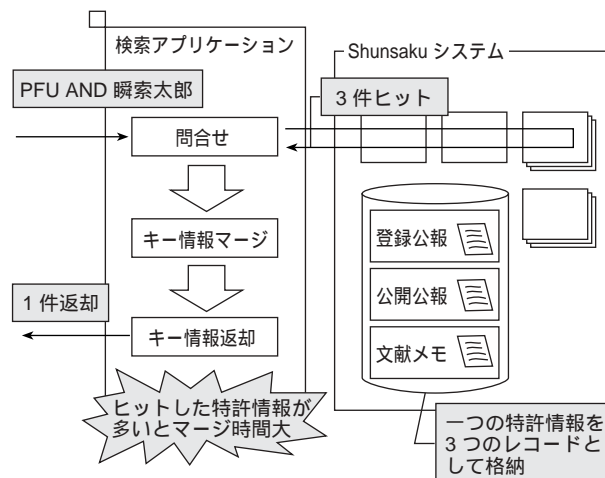


図 4 検索結果マージ処理
(Fig.4-Merge procedure of search results)

る性能の劣化が予想された。

(2) サーバの蓄積データ量の偏りによる検索性能の劣化

Shunsaku システムでの検索性能は、一つのサーバ上に蓄積されるデータ量によって決定される。つまり、サーバを増やし、1 サーバあたりのデータ量を減らすことにより、検索性能は向上することになる（「シンプルなスケールアップ運用」）。

しかし、サーバ間でデータ量に偏りが発生すると、性能が悪いサーバ（管理データ量が多いサーバ）が原因となり性能が悪化する（図 - 5 参照）。

(3) ディスク入出力の発生による検索性能の低下

Shunsaku システムでは、検索用データはサーバ上のメモリに格納されるが、返却用の実データはディレクタサーバのディスク上に格納される。

本システムでは、審査官による一つの検索要求につき最大で 10 万件のキー情報を返す必要がある。しかし、キー情報を実データに埋め込んだ場合、Shunsaku システム上で多量のディスク入出力が発生してしまい、性能劣化を招いてしまう。また、取得した情報（最大 6 MB）のサーバ間転送によるネットワーク負荷も予想された（図 - 6 参照）。

(4) 検索条件過多による検索性能の低下

本システムでは、機能要件として、前回行った検索の結果を再利用し、今回の検索条件と論理積 AND をとって検索を実行する要件がある。この要件を満たすためには、前回の検索結果の保存・再利用が必要である。

しかし、単純に、検索結果の再利用の方式として、検索結果を保存し、AND 条件として今回の検索式に付

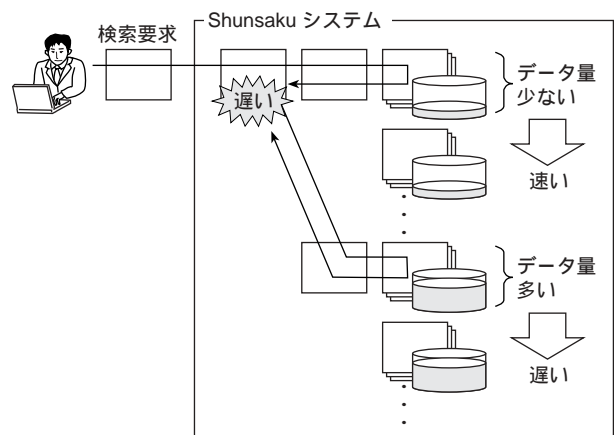


図 5 データ量の偏りによる検索性能の劣化
(Fig.5-Degradation of search performance due to bias in data quantity)

加する方式を採った場合は、最大で 10 万個の OR 検索を実行することとなり、検索性能への影響は明らかであった（図 - 7 参照）。

(5) データ登録時間の短縮

2 時間で 130 万件の特許情報を登録する要件に対して、限られた資源の中、登録時間を大幅に短縮する必要があった。

本システムでは、この要件を満たす為に、Shunsaku の特徴である並列検索機構（ハイトラフィック技術）に伴う検索待ち時間の有効利用が課題となった。

Shunsaku のハイトラフィック技術とは、複数の検索要求を一回の検索条件としてまとめて実行する技術である。まとめて実行することにより、トランザクション量に関わらず、安定した検索時間を提供している（図 - 8 参照）。

その反面、検索時間が一定なため、どんな検索条件

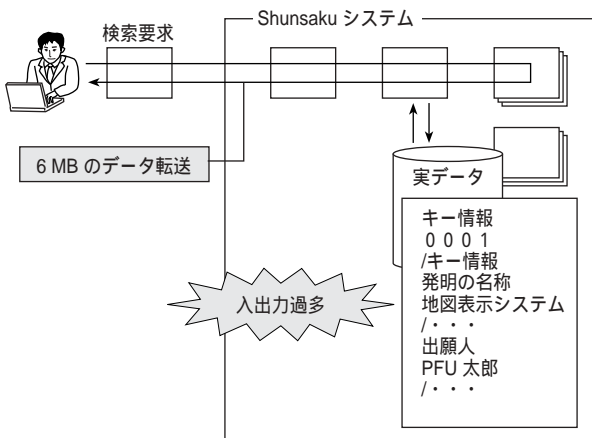


図 6 ディスク入出力の発生による検索性能の劣化 (Fig.6-Degradation of search performance due to excessive tasks of disk input and output procedures)

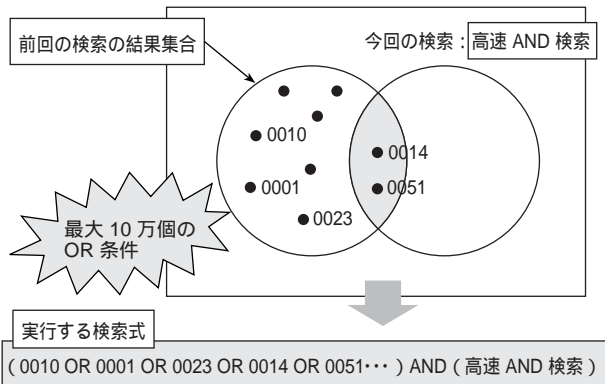


図 7 検索結果集合の再検索 (Fig.7-Re-searching with search result sets)

の検索要求であっても、同じ検索時間（例えば 3 秒）がかかることになる。

データ登録アプリケーションでは、登録済み特許情報を検索し、次いで、検索した特許情報を XML 変換する処理を繰り返す。しかし、単純に処理を多重化し、130 万件を 1 件ずつ処理していたのでは要件を満たすことができない。

上記課題、問題点は以下の(1)~(4)項を開発ポイントとして解決を図った。次節に実現方式の概要を述べる。

(1) 出願情報データの最適化

特許出願情報の XML データを Shunsaku に格納する際に、アプリケーションで扱いやすいように整形して検索効率を向上させる。

(2) XML データ蓄積量の最適化

各検索サーバに蓄積される XML データ量を平均化し、検索性能を最速にする。

(3) ハードウェア資源の有効利用

CPU 利用効率の向上、ネットワーク負荷の低減、ディスク入出力の低減により、検索性能を向上させる。

(4) 検索条件の正規化

Shunsaku に問合せを実行する検索条件を最適化して、検索性能を向上させる。

3.2 実現方式

(1) 特許情報の統合による検索性能の向上

本システムでは、同一特許情報に対する文献データを、一つの XML 文書として統合し、Shunsaku システム上に格納した（図 - 9 参照）。これにより、キー情報のマージを不要とした。

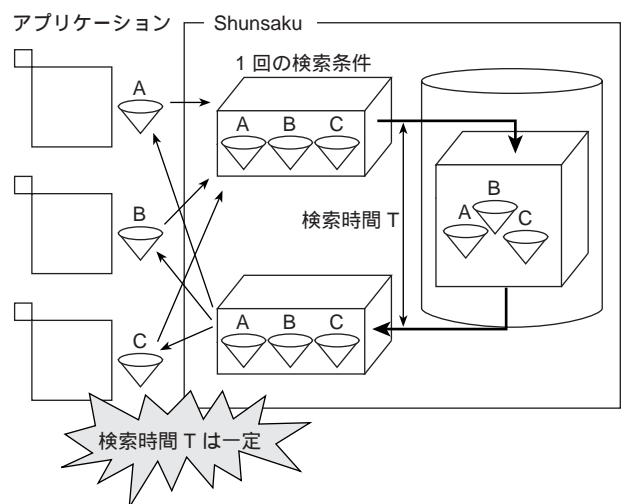


図 8 ハイトラフィック技術 (Fig.8-High traffic technology)

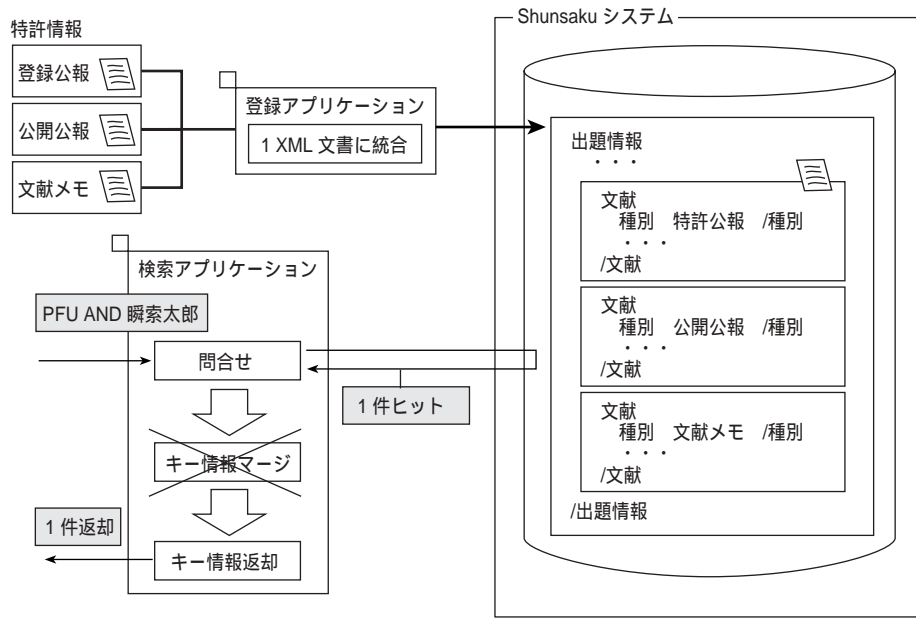


図 9 文献データの統合によるマージ処理の不要化
(Fig.9-Elimination of merge procedure through integration of document data)

(2) 検索データの均等配置による検索性能の安定化

本システムでは、データ蓄積時に、各サーチサーバのデータ容量を監視しながら、一定のデータ容量を超過しないように蓄積し、サーチサーバ間のデータ量に偏りが無いようデータ振り分けを行った(図 - 10 参照)。これにより、検索性能に偏りが生じないようにした。

また、データを均等に配置することにより、予定されているデータ量が増加しても、変わらぬ検索性能を保証している。

(3) 返却情報のメモリ格納による検索性能の向上

本システムでは、Shunsaku システム内部で使用しているレコード ID の空き領域を利用し、検索結果のキー情報をそこに埋め込むことで、大幅に性能を向上させた。

レコード ID は、Shunsaku システム上のメモリに保持されているので、実データをディスク上から読み出す必要がなくなった。つまり、ディスク入出力を発生させず、ネットワーク負荷についても、実データの転送を行わないことで低減することに成功した(図 - 11 参照)。

(4) 検索式保存による検索性能の向上

本システムでは、Shunsaku の特性を活かし、検索結果集合ではなく、検索式を保存する方式とした。検索式保存用の Shunsaku システム(以降、検索式管理 Shunsaku と略す)を構築し、問合せを行った検索式を保存した。

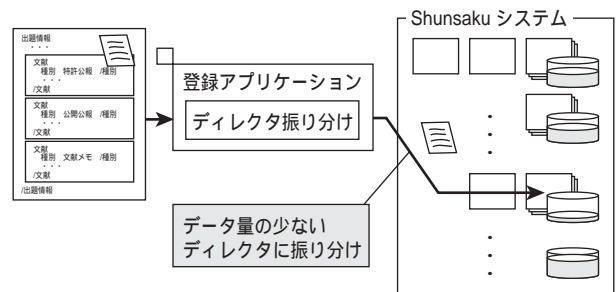


図 10 データ量の均一化
(Fig.10-Homogenization of data quantity)

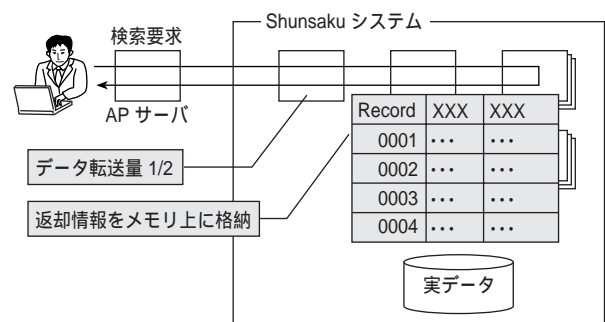


図 11 返却情報のオンメモリ化
(Fig.11-Storing returned information on memory)

検索時、アプリケーションは、審査官から指定された検索式の再利用キー(以降、質問番号)をキー情報として、検索式管理 Shunsaku に問合せを実行し、再利

用する検索式を取得する。取得した検索式を、審査官が
入力した検索式に AND 条件として付加し、文献デー
タの問合せを行う。問合せを行った検索式は、新規質問
番号をキー情報として検索式管理 Shunsaku に保存す
る。検索式の再利用の流れを図 - 12 に示す。

この方式では、検索式を再利用することで、検索結
果である複数の OR 条件が連続することなく検索条件
を絞り込み、検索性能を向上させることに成功した。

また、ユニークキーによるインデックス検索である
再利用検索式の管理に、トランザクション量に性能が依
存してしまうファイルや他の RDB ではなく、文献デー
タと同じ Shunsaku システムを用いることで、検索
式の問合せが性能のボトルネックとなることを防いだ。

(5) Shunsaku の特性を活かした処理の多重化による
データ登録時間の短縮

本システムでは、複数の検索要求をまとめて実行し
ても安定検索ができる Shunsaku のハイトラフィック
技術の特長を活かし、以下の 2 つの対策を実施した。

一つ目は、特許情報を Shunsaku に格納するにあ
たり、1 件ずつ処理するのではなく、一定の件数をまと
めて処理することにより、処理時間の短縮を図った。つ
まり、複数の特許情報のキー情報を OR 条件として検
索し、処理するようにした。

二つ目は、処理を多重化し、片方のスレッドが
Shunsaku の検索待ち時間中に、もう片方のスレッド
が特許情報の XML 変換を行うようにした。つまり、
検索時間 XML 変換処理時間となるよう一度に処理
する件数を調整した(図 - 13 参照)。

これらにより、Shunsaku の特性を活かし、デー
タ登録時間の短縮に成功した。

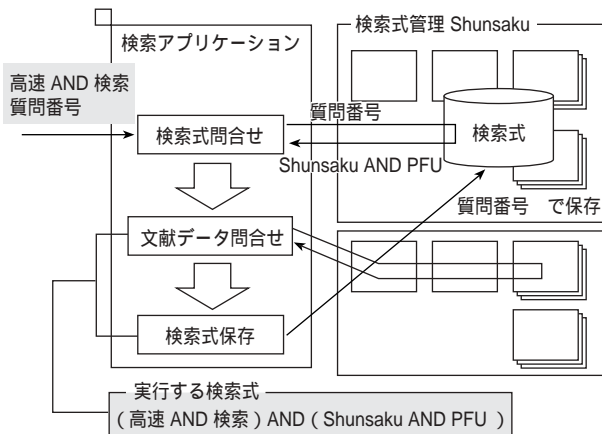


図 12 検索式による再検索
(Fig.12-Re-searching with search statement)

3.3 評価

このような課題に対する取り組みを行った結果、
Shunsaku の特性を活かした検索システムの構築に成
功し、厳しい顧客要件を満足することができた。

(1) オンライン性能要件について

特実全文検索システム構築後に実施した性能評価結
果を表 - 2 に示す。当表に示す通り、ピーク時間帯の性
能要件を満足することができた。

また、図 - 14 はトランザクション量又はヒット件数
と検索時間の関係を従来システムと Shunsaku システ
ムとで対比した図である。これから

- 1) トランザクション量に依存しない検索時間
- 2) ヒット件数に依存しない検索時間

など、Shunsaku の特性を活かしたシステムが完成
されている。

(2) バッチ性能要件について

特実全文検索システム構築後に実施したバッチ性能
評価結果を表 - 3 に示す。当表に示す通り、130 万件
を顧客要件(2 時間)の約 1/2 で処理することに成功
した。

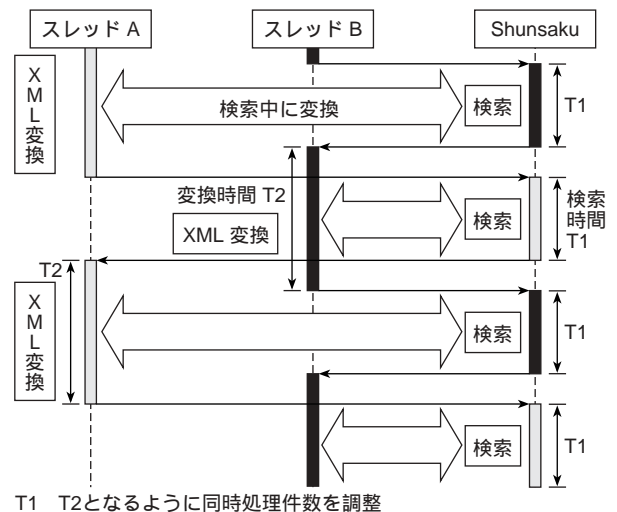


図 13 同時処理件数の調整
(Fig.13-Adjustment of the number of items to be processed
simultaneously)

表 - 2 オンライン性能測定結果

文献種別	ヒット件数		
	10 万件以下	10 万件	10 万件以上
国内文献	3.2 秒	4.0 秒	3.2 秒
外国文献	3.2 秒	-	-
非特許文献	2.1 秒	-	-
WPI 文献	3.8 秒	-	-

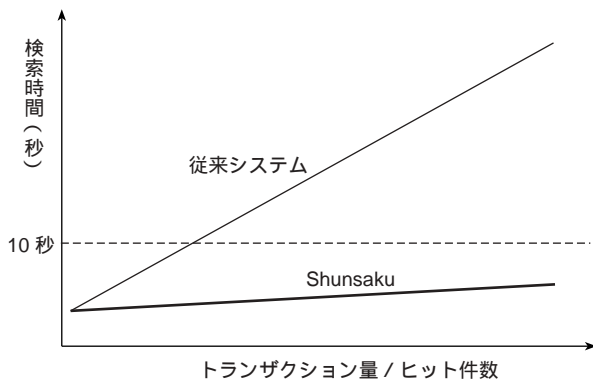


図 14 検索時間とトランザクション量/ヒット件数
(Fig.14-Time required for the search and number of transactions/number of items hit)

また、本性能値は、他の RDB のインデックス作成処理時間と比較しても、高性能であることがわかる。弊社環境における、RDB インデックス作成処理の評価結果を表 - 4 に示す。

4 むすび

これまで PFU の培った Shunsaku 及び、XML 技術を活用することにより、特実全文検索システムとして、厳しい顧客要件を大きく上回る性能を実現した。また、Shunsaku の特長である「インデックスレスの高速検索」、「並列検索機構」、「シンプルなスケールアップ運用」を活かしたシステム構築ができた。

一方、XML 型 DB としての Shunsaku は、非定型なデータを自由に蓄積でき、そのデータを情報として活用できる。これらの特徴を活かすことで、RDBMS に見られるテーブル設計が不要となることから、変化に柔軟に対応可能なシステム構築が可能である。具体的には、異なるシステム間を疎結合する HUB 的なシステムや Web や印刷出版システムで見られるワンソースマルチユースの実現などが挙げられる。

今後は、特実全文検索システムの構築経験や XML 技術の追求により、右記の活動を推進する。

表・3 バッチ性能測定結果

処理件数	処理時間	1件あたりの処理時間
130 万件	52 分	2.4 ミリ秒

表・4 RDB のインデックス作成処理時間

処理件数	処理時間	1件あたりの処理時間
70 万件	60 分	5.1 ミリ秒

(1) Shunsaku のミドルウェアの提供

Shunsaku の特性を活かすファイル管理とデータ管理の機能を持つミドルウェア PKG (PFU 商品名 RichResourceStadium, 以降 RRS と称す) を提供する。これらにより、構築期間の短縮やシステム変更に対応を柔軟な対応を低コストで可能とする。

(2) 業務適用の推進

XML の特長を活かした RRS に業務テンプレートの作成や他システムへの組込みを通じて、業務適用を推進する。その適用業務とすると、以下のシステム適用があげられる。

1) Content Management System

ライフサイクルが短く多様な商品を扱う商品情報 DB。

2) 情報統合 System

SFA, ERP, MES など様々な業務システムが存在しているが、それらを連携可能な HUB 的なシステム。

今後は、Shunsaku を含む XML 技術の中核に、めまぐるしく変化する経営環境に即応可能なシステム構築を目指し、ソリューション提供を行っていく。

参考文献

- 1) Shunsaku 紹介ホームページ
<http://interstage.fujitsu.com/jp/shunsaku/>