オープンソースで実現する
オフィス文書共有システムの構築

FUJITSU ファミリ会関西支部　平成 16 年度

IT フォーラム「Linux 研究会」A グループ

■ 執筆者Ｐｒｏｆｉｌｅ ■

松下　彰良

1988年　株式会社東洋捺染　入社
情報システム室　配属
システム開発担当

浦田　考起

1988年　株式会社コーユービジネス
インフォメーションテクニックス入社
プログラミング，システム開発担当
1995年　株式会社コーユービジネス
システム管理室　システム開発担当
2005年　現在　管理本部　システム部システム課所属
システム開発担当

土屋　和重

1990年　タキイ種苗株式会社 入社
情報システム部所属
システム開発・運用担当

柿本　英治

1990年　株式会社ダイフク入社
　　情報システム部所属

梅田　順也

2002年　ダイヤモンドコンピューターサービス
　　株式会社　入社
　　大阪支店給与人事グループ　配属
　　汎用機での給与計算業務担当
2004 年～現在
　　小型機での給与周辺業務担当

■ 論文要旨 ■

昨年，富士通ファミリ会関西支部主催の IT 研究フォーラムに参加した．フォーラムは半年間に渡り計 11 回開催され，最終回は研究成果発表会が実施される本格的なものであった．さまざまな企業から，年齢も 20～40 才と幅広いメンバーが集い，8 名で構成される我々A グループは，Linux 上で動く全文日本語検索ソフト「Namazu」を元に，オフィスで使われる文書で特に頻度の高い WORD,EXCEL,PDF,HTML を対象にした文書共有システムを構築した．個人で作成された文書を各ファイルサーバにある共有フォルダに置きさえすれば，文書の内容を日本語キーワードで簡単に検索できるものである．これにより，よく似た内容の文書を検索でき，雛型として活用する事で生産性を上げるのが目的である．工夫した点として，1.すべてのソフトをオープンソフトから採用し，費用をかけないでシステムを構築した．2.各部署ごとにファイルサーバが設置されている場合を考慮し，1 度に複数のサーバを検索の対象とする様にした．3.全文検索に必要なインデクスを毎日決められた時間に自動で作る様にし，情報の鮮度を保つ様にした．の 3 点が挙げられる．今回論文として残す事で，今後各社でのシステム構築の基礎にできる様，具体的な事例としてまとめた．

■ 論文目次 ■

■ 図表一覧 ■

１．はじめに

１．１　論文執筆の経緯

FUJITSU ファミリ会関西支部「IT 研究フォーラム」において，『Linux システム勉強会』が開催された．この勉強会では，システム課題としてオープンソフトの日本語全文検索ソフト「Namazu」（なまず）が取り上げられ，参加者はこの「Namazu」を元に各グループで自由に工夫し，独自のシステム構築を通じて Linux の各機能の習得が行なわれた．

我々Ａグループは「Namazu」を何の検索ツールとして用いるのかを検討し，オフィスで最も使われている WORD，EXCEL を全文検索できれば，即戦力となるシステムを開発できると考え，取り組み，それに成功した．その結果をドキュメント化する意味で論文として残す事になった．

１．２　オフィス文書共有システムの必要性

オフィスでは１人に１台のパソコンが用意され，ネットワーク技術の進歩も手伝って，（本社・支社など広い空間を含めた）企業内での人事通達や業績関連資料などの，社内の誰もが閲覧する業務文書，見積書や注文書などの対外的文書，社員個人が業務遂行のために作成したさまざまな資料などは，社内にデジタル化され蓄積しており，その量も質も相当なものであると考えられる．

これらの文書の保存管理は，社内にて一定のルールを策定していなければ，社員個々の保管ルールに依存されるしかないが，実際のところその場合が多いと考えられる．

保管ルールが個々の社員に依存された場合，社内共有文書として扱える膨大な資産でありながら，どんな文書があるのか分からない，探そうにも保管場所すら分からないなど，せっかくの文書情報が実際には共有出来ていないという問題が発生する．

更に保管された文書を探そうとした際，その文書に付けられたファイル名を頼りに探すしか方法がなく，必要な共有文書が保存されているのは分かっていても，検索する事自体に時間がかかってしまうのが普通である．

また，対象ファイルがいくつかに絞れても，文書内容はファイルを開くまで分からないので，目的の文書が見つかるまでファイルを開いては閉じる作業を繰り返さねばならず，無駄に時間を費やす事になりかねない．

そこで，社内に蓄積された EXCEL や WORD 文書の内容全文に対し，キーワードによる検索が行える日本語全文検索ができれば有効なはずだとシステム研究を行った．

このシステムの構築・導入は，共有オフィス文書の検索時間の短縮と，文書共有によるナレッジマネジメントの推進を目的としている．

２．システム

２．１　システム概要

図１．システム概要

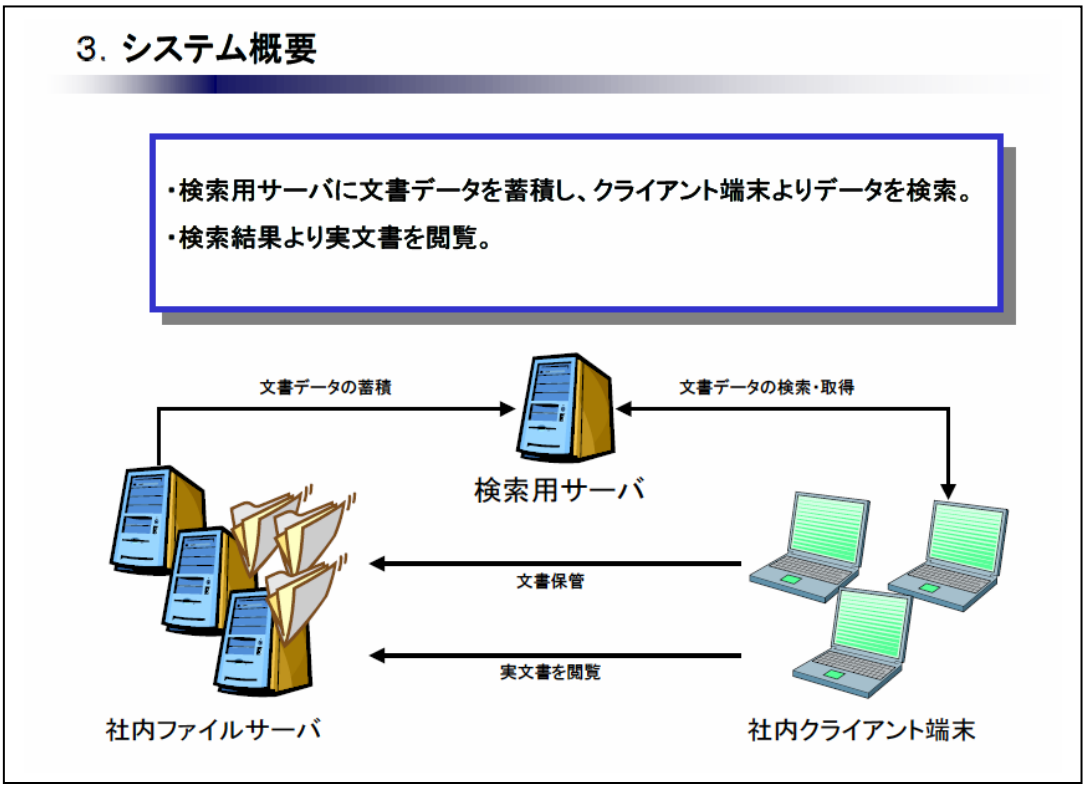


図１は全文検索システムのシステム概要である．

今回検索の対象とするものは，社内のクライアント端末からアクセスできる複数のファイルサーバに保管されている共有文章である．

社内には検索用サーバを Linux で新たに設置する．

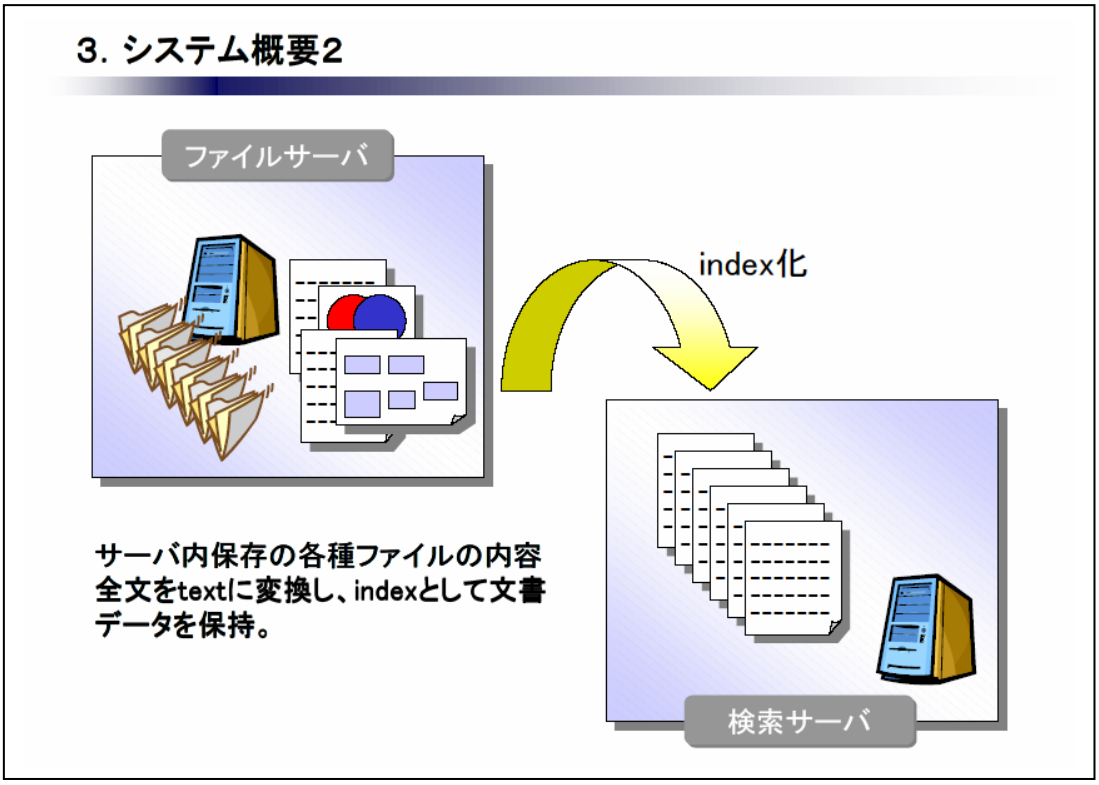
検索用サーバは，ファイルサーバ内の文書に対し，文書に含まれる検索ワードを収集し，それをインデクス・ファイルとして蓄積する．インデクス・ファイルとは検索ワードと被検索のファイルとを紐付ける参照アドレスの様なファイルであり，一般的にＯＳに標準装備されるファイル内の文字列検索と異なるところは，このインデクス・ファイルが前もって準備してある事で，高速な全文検索が実現できる事である．すなわち検索の度にすべてのファイルの中身を検索するのではなく，あらかじめすべての文書を検索し，結果だけをサーバに蓄えておく事で効率化を図っているのである．

各社内クライアントからは，検索用サーバの Web サーバ機能（web ブラウザを使う）から共有文書への検索要求を行い，検索結果を瞬時に受け取る事ができる．web ブラウザ経由で「問い合わせ」と「結果」の参照ができるので，各クライアント端末には特別なソフトウェアをインストールする必要は無い．

各クライアントの結果からは，社内ファイルサーバのどこに自分が必要とするファイルがあるかが分かり，ユーザはそのファイルを即座に閲覧する事が可能となる．

２．２　システム詳細（構成）

図２．システム概要２



次に，検索用サーバに蓄積される文書情報について説明する．（図２．システム概要２に記載）

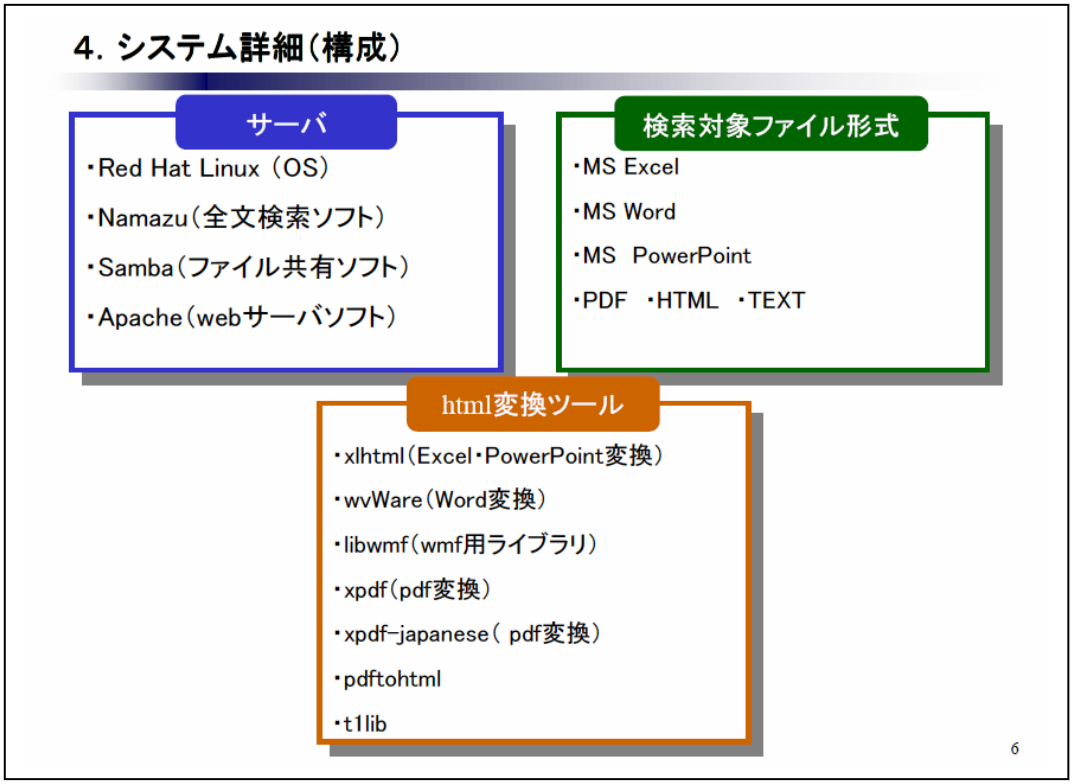
ファイルサーバ内には，様々なファイル形式で共有文書が保管されている．そこで一般的にオフィス内でよく使用される「EXCEL」「WORD」「Power Point」「PDF」「HTML」そして「TEXT」ファイルの文書を検索の対象にした．

これらのいろいろな形式の文書は，文書内に含まれる単語をそのファイルと関連付けるためのインデクスとして検索用サーバに保管するため，その形式を一旦プレーンなHTML（TEXT）形式にそれぞれ変換する事になる．変換するためのツールは全てオープンソフトがインターネット上にあり，検索用サーバにインストールすれば「Namazu」が自動的にリンクをしてくれる．

また文章を単語化するために，分かち書きのツールでやはりオープンソフトの「KAKASI」をインストールしておく．日本語の分かち書きについては英語のそれと異なり，連続した文字列の中から意味のある単語を抽出しなければならない．「KAKASI」には標準で付属している辞書に約 10 万語を超える単語が含まれており，また個人名や専門用語等，辞書に追加登録もできるので，十分に実用的な品質を備えている．「KAKASI」により，ファイルサーバ上のすべての文書を検索文字列としてインデクス化する事ができる．

すなわち，［ファイルサーバ上の文書をプレーンテキストに置き換える］→［分かち書きを解析する］→［インデクス・ファイルを作成する］→［インデクス・ファイルを元に検索を行う］という流れを汲む事で，様々な形式の共有文書に対して，高速な日本語全文検索を可能にしている．

図３．システム詳細（構成）



次に，システムを構成しているオープンソースについて記述する．（図３．システム詳細（構成）に記載）

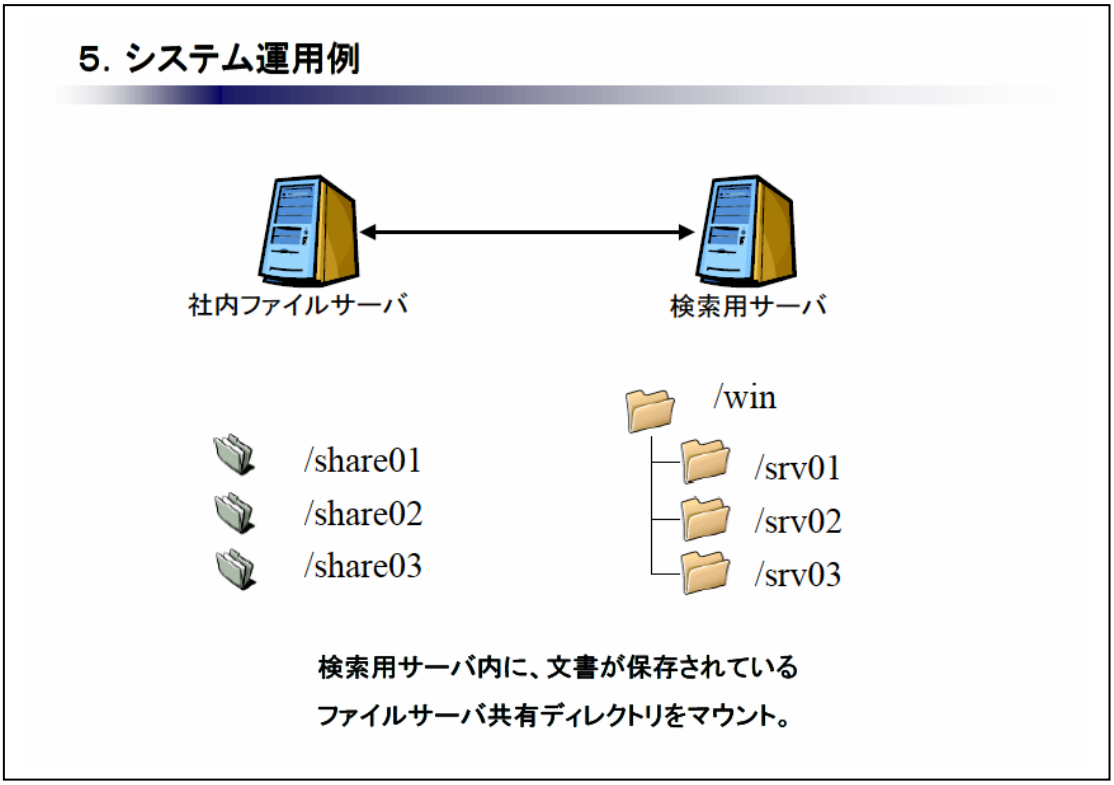
検索サーバには OS として「Red Hat Linux」，日本語全文検索ソフトとして「Namazu」，Windows 系ファイルサーバとのディレクトリ共有に必要な「Samba」，クライアント端末からブラウザによる検索実行のために「Apache」を導入している．

また，各種ファイル形式の HTML への変換ソフトでは，EXCEL と PowerPoint の変換には「xlhtml」，WORD 変換に「wvWare」「libwmf」，PDF 変換に「pdftohtml」「xpdf」「xpdf-japanese」「t1lib」をサーバに組み込んでいる．

これらのソフトウェアはすべてがオープンソース（無料）であり，インターネットから簡単に入手する事ができる．

２．３　システム詳細（設定）

図４．システム詳細（設定 1）

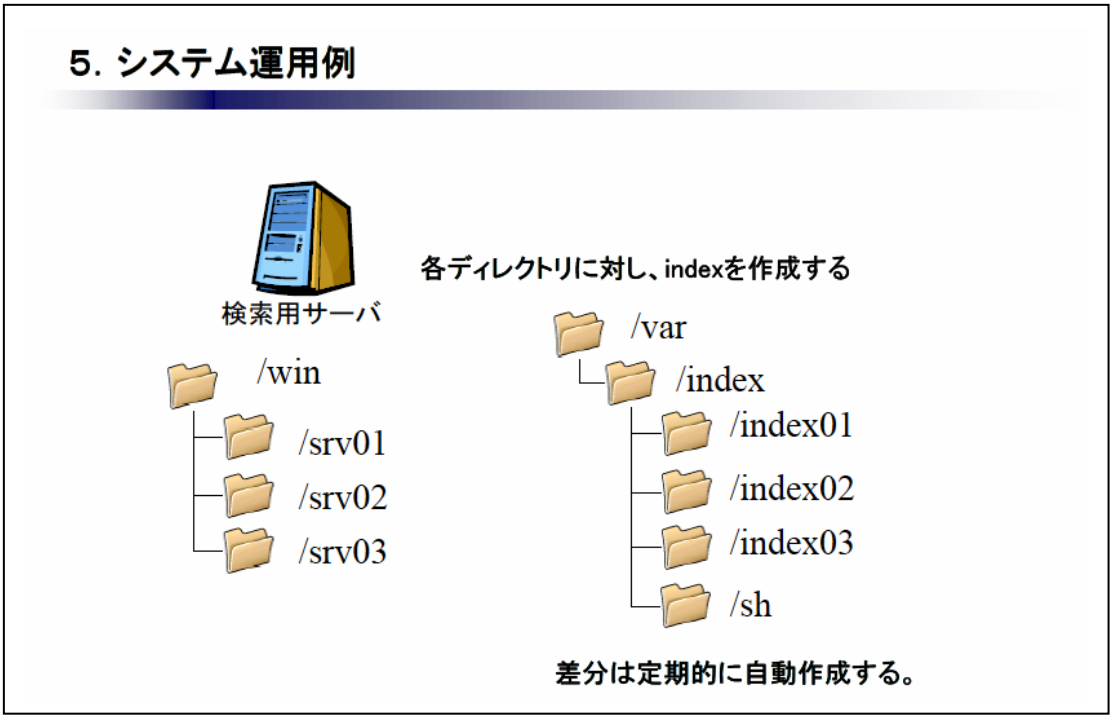


検索サーバに Linux サーバを使用しているが，ファイルサーバには実際のオフィスでの使用環境を考慮して Windows のファイルサーバを使用した．

図４システム詳細（設定 1）の記述にある　share01～share03 までをそれぞれ別々の３台のサーバと見立て，検索用サーバ/srv01～/srv03 のそれぞれに「Samba」を使ってマウントする．この作業によって，検索用サーバからは自分自身のサーバ内に各ファイルサーバ内のファイル全てが存在するかの様に扱う事ができる．

「Namazu」の実行プログラムは基本的に，自サーバ内の選択したディレクトリに対して，それ以下に存在するファイルをインデクス化する．そのためこの例の様に，散らばったファイルサーバの共有ディレクトリを仮想的に検索サーバ自身のマシン内にあるディレクトリの様に見せなければならない．「Samba」で各ファイルサーバの共有ディレクトリをマウントする事によって「Namazu」の実行プログラムは，全ファイルサーバのファイルを検索対象とする事ができる．

図５．システム詳細（設定 2）



検索用サーバの文書ファイルにアクセスが可能になった．次はインデクス・ファイルの作成について記述する．（図５．システム詳細（設定 2）に記載）

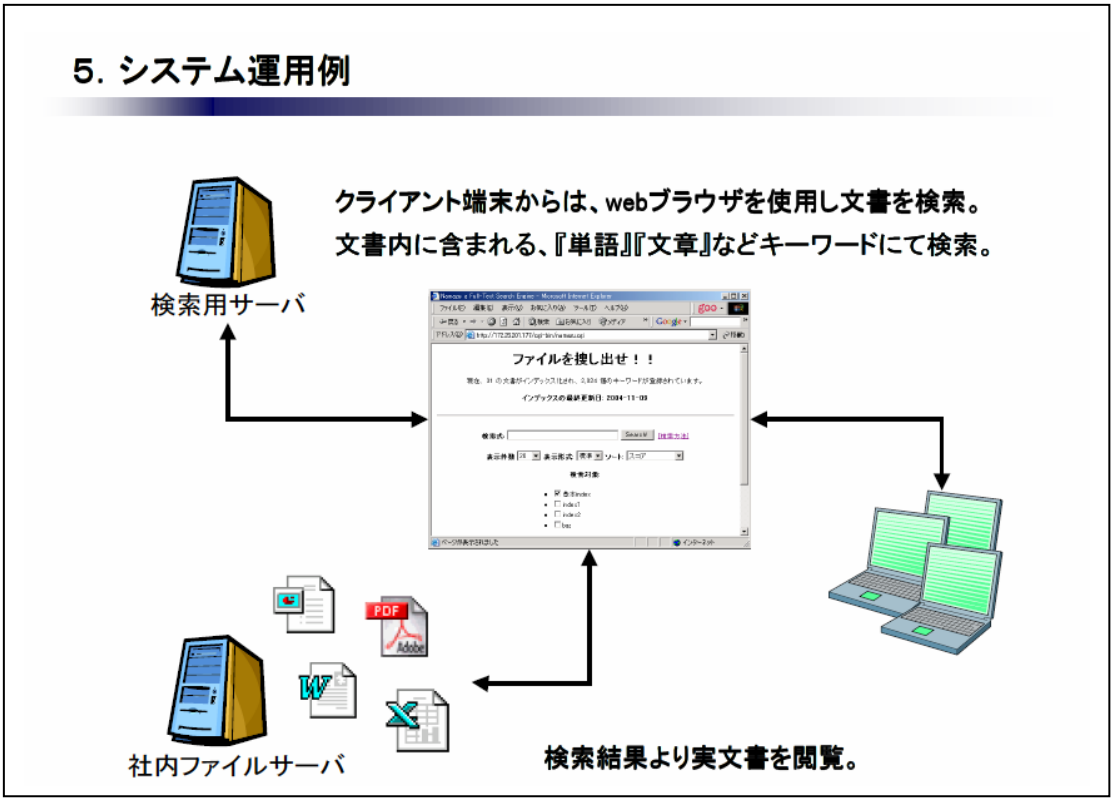
検索を実行する際，インデクス化されて高速な検索ができるとはいえ，すべてを検索するのでは多少の無駄が生じてしまう．おおよそどのサーバに自分が欲している文書が保存されているかが分かる場合は，そのサーバ内だけの検索で済ませる事ができれば，より検索時間の短縮につながる．

そこで各検索対象サーバごとに/index01～/index03 というインデクス格納用のディレクトリを作成して，そこにインデクス・ファイルを格納する．これによってブラウザから検索を実行する際に，検索の単位となるグルーピングができる．

/sh の中にはインデクス作成用のシェルスクリプトを設置し，srv01→/index01 ，serv02→/index02 ，srv03→index03 という様にインデクス・ファイルを作成する．このインデクス・ファイルを作成するシェルスクリプトを「Linux」で標準的に利用されているスケジューリング機能である「cron」を利用し，毎時実行する事によって，日々社内にて作成され，増加する文書を自動的にインデクス化していく．

基本的に初回のインデクス作成以降は，差分のインデクス・ファイルを作成する事ができるので，処理時間とサーバ負荷を最小にする事ができる．サーバのスペックに余裕がある場合は，狭い間隔で実行する事で，検索対象の文書と実際に保存されている文書のギャップを埋める事が可能となる．実際には１日１回，昼休みの時間帯か，アフターファイブの時間帯に行う事になると思われる．

図６．システム詳細（運用イメージ）



最後にエンドユーザの使用方法だが，検索の際には，個々の社員は各クライアント端末の Web ブラウザを使用し，文書内容全体に対しキーワード検索を実行する．ファイルサーバ毎に分けられたインデクスを検索する事によって社内ファイルサーバにある各種ファイルを検索される．

検索結果から必要な文書が見つかれば，そのリンクをクリックする事で実文書を閲覧する事ができる．直接編集を行う事はできないが，検索した文字列が含まれるファイル名，保存されているサーバ名，ディレクトリ名の情報を得る事ができるので，その情報から実ファイルにアクセスする事が可能となる．（図６．システム詳細（運用イメージ）に記載）

３．性能の検証

３．１　インデクスの作成時間

検索対象ファイルとして，ページ数の多い PDF ファイルを複数コピーし，インデックスを作成した．検索対象のファイル数や頁数とインデクス作成に要する時間を，表１．インデックス作成時間実測にまとめた．インデクス自体の作成は比較的速く，ページ数３倍に対して作成時間 1.6 倍と，インデクスのディスク書込みに時間を要していると思われる．これによって１時間でファイル 1,200 個分（約 51,600 頁分）のインデックスが作成できる事が分かる．差分を定期的に自動作成する運用で，十分運用に耐えうる事が分かる．

表１．インデックス作成時間実測

ファイル数	頁数	作成時間（sec）
1	43	9
2	86	12
3	129	15

３．２　インデクスの必要容量と検索時間

検索対象ファイルのインデクス容量調査の結果を表２にまとめた．インデクスは文字データのみ抽出するため，サイズは元ファイルの数%になる．

また，インデクスの容量が 158MB となった文書を検索に要する時間を測定すると，1 秒前後であった．検索そのものについても，ほとんどストレスなく利用できる事が分かる．

表２．インデックス容量実測

対象ファイル	インデックス対象ファイルサイズ(KB)	インデックスサイズ(KB)	圧縮率
doc	195.0	18.3	9%
xls	109.5	46.7	43%
ppt	3572.0	51.0	1%
pdf	12352.9	240.3	2%

３．３　導入効果と導入費用

導入効果については，いろいろな使われ方ができるシステムであると自負しているので検証が難しいが，例えば EXCEL で作成されているファイルレイアウト群から，ある項目が使われている全てのファイルを探し出したい時，また WORD で作成された仕様書群から，あるキーワードが使われているものを全て探す時など，効果の高い使われ方が期待できる．また導入費用は一切かかっていない．開発期間は概算１人月くらいである．（８名×４回×５時間で計算）

４．注意事項と今後の課題

(1) インデクス作成時のスケジュールには，検索サーバとファイルサーバの電源管理を考慮する必用がある．
(2) 検索結果ファイルの更新権限に配慮する必要がある．基本的にはコピーして使う事を徹底しなければならない．
(3) ファイルのセキュリティに気を配る必要がある．文書によっては取り扱いのできる人または部署が限定される場合がある．今回はこの点について全く考慮できなかった．
(4) VB 等プログラム・ソースを全文検索できれば，ある命令が使われている AP や，あるファイル項目が使われている AP が検索できるので更に便利になるはずである，続けて研究して行きたい．

５．おわりに

今回の「IT 研究フォーラム」の活動では，コストを全くかけずに Windows 環境下の日本語文書全文検索システムの構築を完成する事ができた．また，Linux をはじめとするオープンソースのメリットを実感する機会となり，研究会で構築するシステムは各メンバーの社内で実際に適応する事を前提としたので，より実践的な研究がはかれた．更に，本論文の作成によって，半年間の研究成果をまとめる事ができ，今後 Linux の導入を検討，開発しようとする方々の参考になるものと確信する．

最後に「IT 研究フォーラム」でご指導して頂いた FUJITSU ファミリ会幹事の谷様・森田様，アドバイザーの富士通四国システムズの高科様・武田様，関西支部事務局の尾初瀬様のご支援に対し心より御礼を申し上げます．

参考文献

［１］10日でおぼえるRed Hat Linux 9サーバ構築・管理入門教室
松本光春著，翔泳社（2003年７月発行）
［２］Namazuシステムの構築と活用～日本語全文検索徹底ガイド～
馬場　肇著，ソフトバンク・パブリッシング社（2003年７月発行）