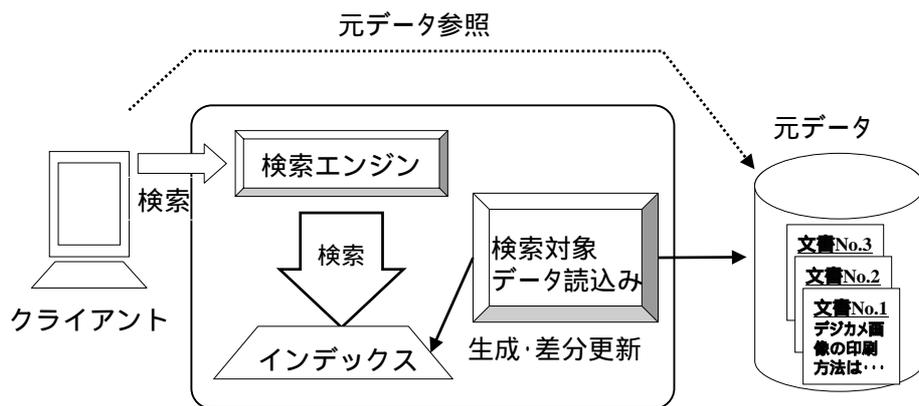


エンタープライズサーチ ～最新の技術動向と活用事例～

2007年11月19日
住友電工情報システム株式会社
伊藤 彰康

1

情報検索システム概念図



2

情報検索システムの性能

- ▣ 検索精度
 - 以下の3つの要素を持つテスト・コレクションを用いて評価
 - ・検索対象となる文書集合
 - ・検索質問集合
 - ・各検索質問に対して適合する文書集合の対応表
- ▣ 検索速度
- ▣ インデックス生成 / 差分更新速度
- ▣ 検索容量

3

検索精度:再現率と適合率

- ▣ 文書集合に対する検索質問の結果

| | 検索された文書 | 検索されなかった文書 |
|-------|---------|------------|
| 適合文書 | w | x |
| 非適合文書 | y | z |

- ▣ 再現率 (recall) = $w / (w + x)$
検索結果にどれだけ漏れがないか
- ▣ 適合率 (precision) = $w / (w + y)$
検索結果にどれだけノイズがないか

4

企業内システムの検索機能に対するユーザーの声

検索が項目単位でしかできない

大文字小文字を区別して検索しなければならない

ほしい情報がどこにあるかわからない

検索対象が散らばっていて画一的に検索したい

レスポンスが悪い



5

企業内に散らばる検索対象

Word/Excel/PPT
PDF/テキスト等

部門
ファイルサーバ

部門技術共有システム

検索対象がいろいろな場所・システム・形態で散らばっており、探したい情報が瞬時に引き出せない

Word/Excel/PPT
PDF/テキスト等

全社
ファイルサーバ

製品設計システム

全社人事システム

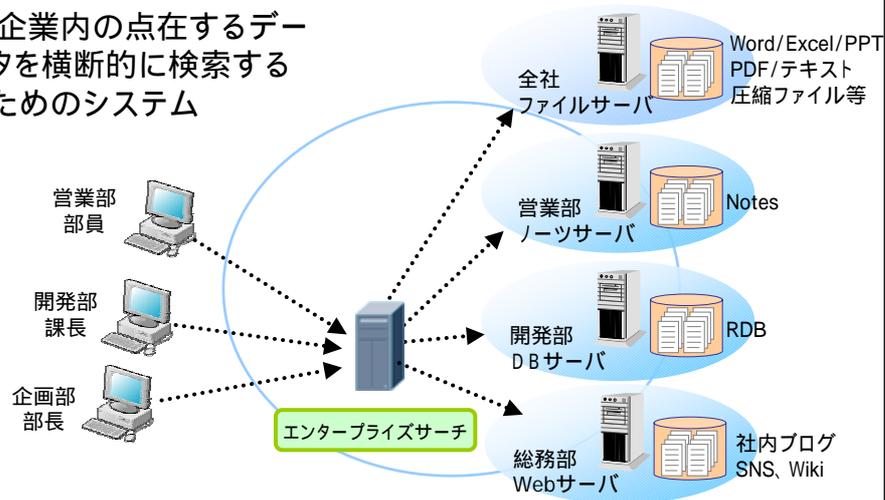
どこをどうやって調べればいいんだろう？



6

エンタープライズサーチとは

企業内の点在するデータを横断的に検索するためのシステム



7

エンタープライズサーチの要件

- 検索結果に対する要件
 - 欲しいものを速く見つけたい
 - ハイパーリンクに頼らない検索ランキング
- 検索結果以外の要件
 - データ容量およびアクセス増に対する拡張性
 - 連続運用が容易
 - アクセス権限管理、セキュリティ対策
 - 多彩なデータ形式に対応
 - 国際化対応
 - カスタマイズ

8

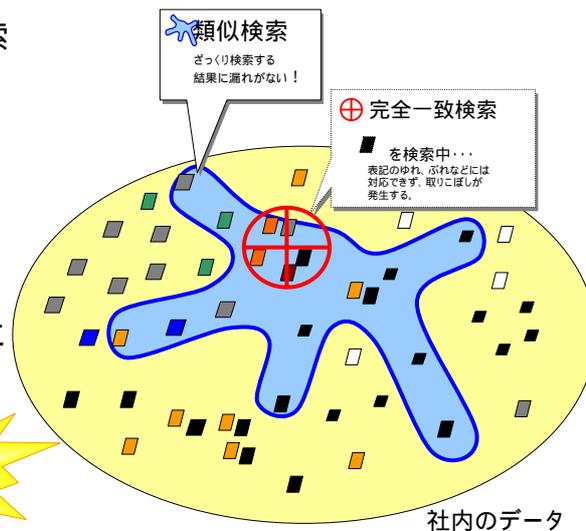
検索ランキング技術

- 全文検索アルゴリズム
 - 索引語の重み付け
 - N-gram vs. 形態素解析、構文解析
 - 類似検索 (自然文検索、あいまい検索、表記揺れ対応、もしかして)
 - 位置考慮検索 / 近傍検索
- 検索質問拡張
 - キーワード抽出・関連語抽出による気付き支援、関連検索、サジェスト
 - 関連語辞書、シソーラス連携
 - シソーラス自動構築
- 検索結果の分析 / 分類 / 可視化
- 組合せ検索、多重ソート、カテゴリ分類 **メタデータ(属性)を活用**
- アクセスランキング、レコメンデーション **アクティビティログを活用**
- リンク解析: GoogleのPageRank **ハイパーリンクを活用**
 - 多くの価値のあるページからリンクされているページは価値のある重要なページ
ファイルサーバ、DBサーバ等の検索では使用できない

9

エンタープライズサーチの検索精度

- 企業内検索では検索漏れが致命的になる場合が多い
- 検索漏れを防ぐ
- 検索結果上位のノイズを削減する
- 検索結果を効率的に絞り込む



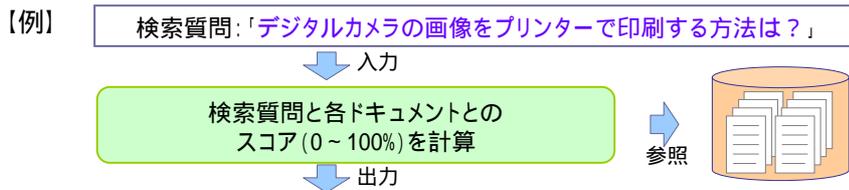
10

N-gram、形態素解析、ハイブリッド方式の比較

| | N-gram方式 | 形態素解析方式 | ハイブリッド方式 |
|----------|--|--|---|
| システム辞書 | 不要 | 必要 | 必要(必須ではない) |
| インデックス形式 | N-gram転置ファイル :フルテキストインデックス | 単語転置ファイル :単語インデックス | N-gram転置ファイル + 単語転置ファイル |
| 長所 | 任意の文字列を検索可能 検索漏れがない | 検索結果にノイズが入り難い | ・任意の文字列を検索可能 ・検索漏れがない ・検索結果(上位)に ノイズが入り難い |
| 短所 | 検索結果にノイズが入り易い ・「京都」で検索して 「東京都」がヒット。 ・「スキー」で検索して 「ウイスキー」がヒット。 | 検索漏れが生じる リカバーできず致命的 新語・造語が検索できない 辞書のメンテが必要 インデックス再生成が必要 ・型番の部分一致 ・「イントラ」で検索して 「イントラネット」がヒットしない。 | ・インデックスサイズが大きい (従来比1.1倍程度) ・インデックス更新時間は従 来と同程度(ただし、CPU、メ モリの負荷は高い) 検索漏れはN-gramベース でカバーできるため辞書のメ ンテおよびインデックスの再 生成は必須ではない |

11

類似検索



| スコア | 検索結果 |
|-----|-------------------------------|
| 89% | デジタルカメラ画像をプリンターで印刷するには... |
| 76% | デジカメ画像をプリンターで印刷する場合は... |
| 67% | デジタルカメラの写真をプリンタで出力する方法について... |
| 50% | プリンタで写真を印刷するには... |
| 40% | デジタルカメラの使用法について教えます... |

「デジタルカメラ」と「デジカメ」、「プリンター」と「プリンタ」等、表記の揺れを吸収。

- キーワードやフレーズでも類似検索が可能
- キーワード全文検索を圧倒する回答力を実現

12

類似検索アルゴリズム

ステップ1：部分文字列の選別

検索質問から出現頻度をもとに検索に有効な部分文字列を高速選別

- 検索精度を維持して検索処理を高速化
- 検索質問が長文であっても大丈夫

(例) 検索質問: 「デジタルカメラの画像をプリンターで印刷する方法は？」
 「デジタルカメラ」「画像」「プリンター」「印刷」
 + 「デジ」「ジタ」「タル」「ルカ」「カメ」「メラ」...

インデックス

文書 1
 「デジカメ」の機能は、...
 「デジカメ」の印刷機能は、...

ステップ2：スコアの算出

各部分文字列の出現頻度および出現集中度を考慮した重みを加算

- 出現頻度: 小(出現する文書数は限られる)
- 出現集中度: 大(同じ文書に繰り返し出現)

(例) 文書1のスコア
 = 「デジ」 + 「カメ」の重み(2回出現を考慮) + 「印刷」の重み

13

類似検索アルゴリズム(続き)

辞書が不要なN-gram方式ベースにより検索漏れを防止
 どのような分野のデータでもOK

独自の統計処理による検索結果のランキング

$tf \cdot idf$ + 出現集中度を考慮したスコアの算出

N-gramベース + 形態素解析ベースの類似検索

表記の揺れを吸収する漏れのない類似検索

+ キーワードの有無をより重視した類似検索



検索漏れを防ぐ(再現率向上)

検索結果上位にはノイズが少ない(適合率向上)

14

位置考慮検索

- ある程度の限られた範囲(先頭)に、検索質問で指定した文章のキーワードが集中
正解である確率が高い
 - 限られた範囲
数百文字
 - 検索質問で指定した文章の全てのキーワードが含まれていても、キーワードが点在している
不正解である確率が高い
- ↓
- 指定した数百文字の範囲におけるキーワードの密度を考慮して、スコアを計算し、検索結果の順序付けを行う



15

類似検索(あいまい検索)と検索質問拡張

- 類似検索のみ
- 表記の揺れを吸収できる
 - 組み合わせ ↔ 組合せ
 - スペルミスカバーできる
 - インターネット ↔ インタネット
 - 短縮表現に対応できる
 - デジタルカメラ ↔ デジカメ

検索質問拡張

 キーワード抽出
関連語抽出
(気付き支援)

 関連語辞書
(ユーザ登録)

 シソーラス連携
(IT、企業名、医学
等、分野ごとの辞書)

 画面 モニタ
PC パソコン
ANA 全日空

16

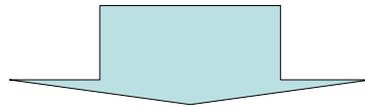
シソーラス自動構築

■ 関連語辞書

- 管理者が用語登録する必要がある
- 社内検索に必要な最小限の用語を登録可能

■ 市販シソーラス

- 用語登録不要
- 社内検索に不要な用語が登録されている可能性がある
- 高価



■ シソーラス自動構築

- 社内データから独自の用語を自動抽出
- 実際に使用されている用語から構成される検索に有効なシソーラス

17

シソーラス構築手順

- 1 文脈列の抽出
 - 2 順序対の抽出
 - 3 候補対の選出
 - 4 関連語判定
- 似たような扱われ方をしている単語は似ている
 - 前後ワードおよび文字に共通するものがある単語
 - 前後ワードおよび文字の異なり数もカウント

[原稿]を[プリント]して[郵送]する

[原稿]を[印刷]して[郵送]する



自動抽出された関連語対

| 関連語 1 | 関連語 2 |
|----------|-----------|
| SCSI | USB |
| スクリーンサーバ | スクリーンサーバー |
| 横書 | 縦書 |
| ディスプレイ | モニタ |
| USB | スキャナ |
| 文書 | 文章 |
| スキャナ | デジカメ |
| スクリーンサーバ | 壁紙 |
| SCSI | スキャナ |
| アドバイス | 質問 |
| BMP | JPEG |
| オフライン | 切断 |
| キーボード | ケーブル |
| SCSIカード | SCSI接続 |
| 全角 | 半角 |
| アップグレード | バージョンアップ |
| USB | ケーブル |
| EXE | exe |
| デジカメ | デジタルカメラ |
| テキストボックス | 挿入 |

18

組合せ検索

- 全文検索
 - 類似検索 : 自然文によるあいまい検索
 - キーワード検索 : キーワードを指定して完全一致検索
- 属性検索 : 数値、日付、カテゴリ等のフィールド検索

19

カテゴリ分類

20

エンタープライズサーチの要件

- 検索結果に対する要件
 - 欲しいものを速く見つけたい
 - ハイパーリンクに頼らない検索ランキング
- **検索結果以外の要件**
 - データ容量およびアクセス増に対する拡張性
 - 連続運用が容易
 - アクセス権限管理、セキュリティ対策
 - 多彩なデータ形式に対応
 - 国際化対応
 - カスタマイズ

21

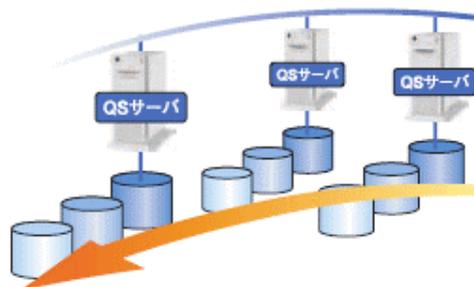
データ容量に対する拡張性

数テラバイトのファイルサーバの検索にはサーバが何台必要？

- 1台の検索サーバでの検索容量
- 1台の検索サーバの検索容量を超えたときの拡張性

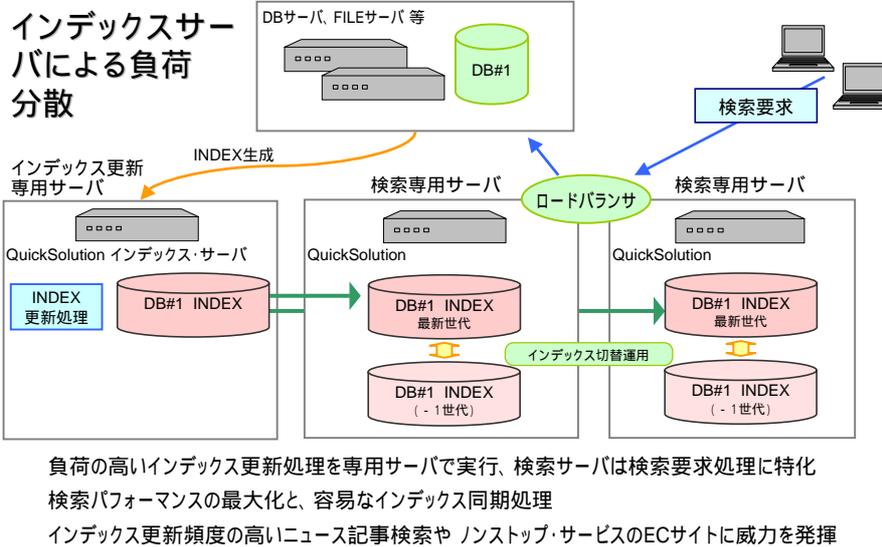
分散検索

■分散検索



22

アクセス増に対する拡張性



23

高い処理性能や大容量データへの対応

■ OSが64ビットでもアプリケーションは？

インデックス生成時間の短縮など64bit化により更に高機能化ができます。

大容量のデータを扱い、高度な処理を求められる検索エンジンは、高い信頼性を保つためにもメモリを確保したい。

64ビットOSは4GBを超えるメモリの利用が可能

+

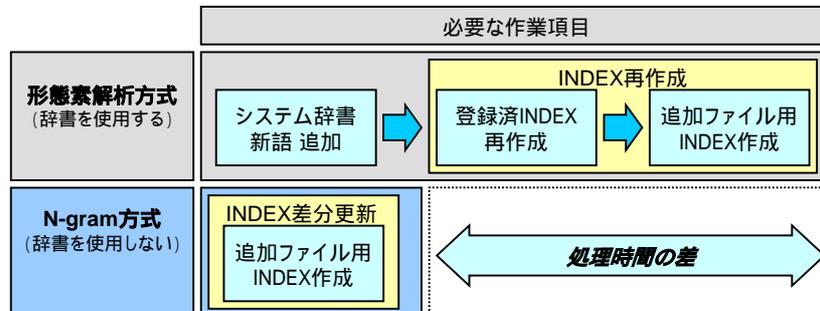
64ビット対応アプリケーション

大規模向けサーバを使用した高信頼性のシステムを構築可能

24

連続運用: N-gram方式は保守運用が容易

「新たな検索語句の追加」 + 「検索対象ファイルの追加」が発生した場合、その検索語句を用いて、任意のファイルの検索を行うためには...



形態素解析方式の場合、

新たな検索語句を事前にシステム辞書に登録した後、追加ファイルだけでなく、元のファイル全てに対してINDEXの作り直しが必要。

N-gram方式 ならば、

追加されたファイルの分だけ、INDEXを追加(差分更新)するだけでOK!!

N-gram方式でも、インデックス最適化のため定期的にインデックス再生成が必要な製品も多い

25

アクセス権限管理、セキュリティ対策

- ユーザ認証
 - 独自認証、Active Directory連携、LDAP連携、シングルサインオン
- アクセス制御
 - 検索対象DBごとのアクセス制御
 - 文書単位の権限継承
 - リアルタイム権限継承 vs. ACLのインデックス取込
- 検索ログ収集・集計・レポート
 - 検索ログ収集(だれが、いつ、何を、どのようなキーワードで検索したか)
 - 収集したログの集計・レポート
 - この1週間で使われた検索キーワード一覧等

26

ユーザ認証とアクセス制御

ユーザとグループ

| | | |
|----|--------|--------|
| A氏 | 営業グループ | |
| B氏 | 開発グループ | 企画グループ |

- ・独自認証
- ・Active Directory連携
- ・LDAP連携

グループ毎のアクセス制御

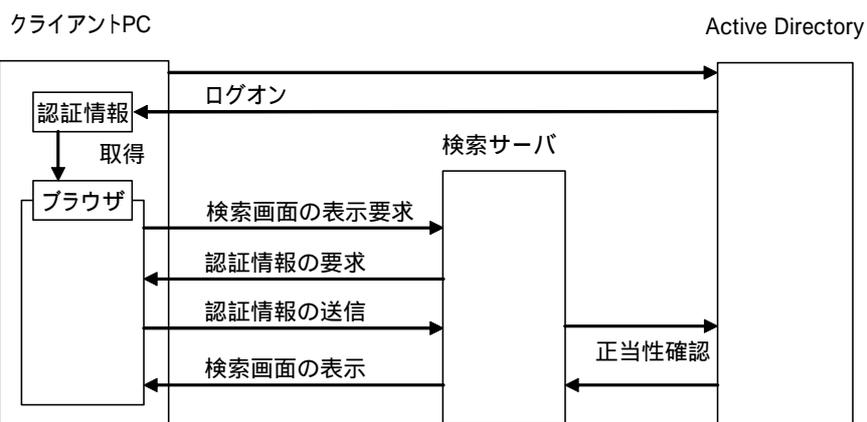
| | 営業グループ | 開発グループ | 企画グループ |
|-------------|--------|--------|--------|
| 営業CSVデータ | 検索可 | 検索不可 | 検索可 |
| 開発Oracleデータ | 検索不可 | 検索可 | 検索不可 |
| 営業プレゼン資料 | 検索可 | 検索不可 | 検索不可 |
| 開発XMLデータ | 検索不可 | 検索可 | 検索不可 |
| 経営データ | 検索不可 | 検索不可 | 検索可 |

結果

| | A氏 | B氏 |
|-------------|--------------------------|--------------------------|
| 営業CSVデータ | <input type="checkbox"/> | <input type="checkbox"/> |
| 開発Oracleデータ | × | <input type="checkbox"/> |
| 営業プレゼン資料 | <input type="checkbox"/> | × |
| 開発XMLデータ | × | <input type="checkbox"/> |
| 経営データ | × | <input type="checkbox"/> |

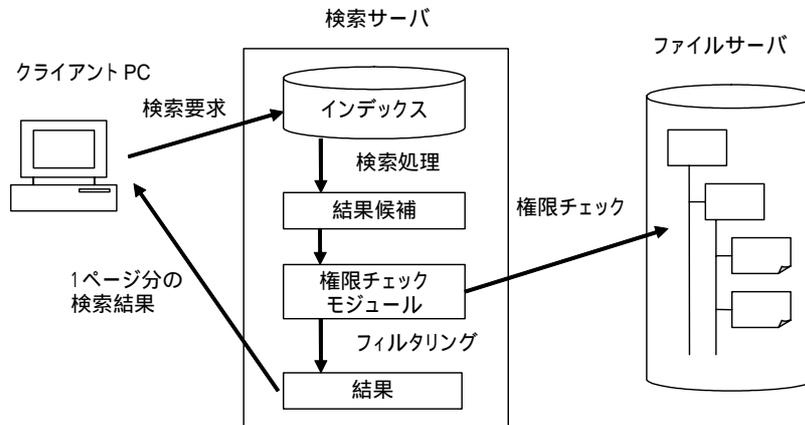
27

Windowsログオンによるシングルサインオン



28

権限継承の仕組み



29

多彩なデータ形式に対応

- データベースの検索
 - RDBを利用して構築されている基幹システムの業務データの検索
- 動的生成ページのWebクロールング
 - 社内ブログ、SNS、Wiki等
- メタデータの検索
 - RDB、XML-DB / XMLファイル、文書ファイル、画像ファイル、Webページ
 - 組合せ検索
 - 項目ごとの属性検索
 - 項目をまたがる全文検索
 - 特定項目を重視した検索等

30

RDBの限界を打ち破るエンタープライズサーチ

通常のRDB検索 (SQL)

9月の請求で、請求先が「楽々～」、
金額が1,000,000円以上

SQL

範囲指定 → 部分一致 → 範囲指定

| 請求 | 請求日 | 請求先 | 請求件名 | 請求金額 |
|----|------|--------|-----------|-----------|
| | 8/30 | A銀行 | 支店LAN工事 | 3,800,000 |
| | 9/6 | B自動車 | 9月度保守費 | 8,250,000 |
| | 9/8 | 楽々電気工業 | コンサルティング費 | 1,350,000 |
| | 9/11 | 楽々商事 | 楽々電工向けPC | 450,000 |

| 受注 | 受注日 | 顧客名 | 受注件名 | 受注金額 |
|----|------|--------|-----------|------------|
| | 8/26 | C運輸 | 新システム機器一式 | 12,500,000 |
| | 9/4 | 楽々電気工業 | 9月度保守費 | 6,000,000 |
| | 9/5 | システム開発 | 楽々電工保守代行 | 2,500,000 |

- 項目を特定した固定的な検索
- 部分一致検索は応答が遅い

さらにエンタープライズサーチでは...

9月の請求で、「楽々電気工業」関連
「楽々電工」関連の全請求と受注

属性検索 → 類似検索 → 類似検索

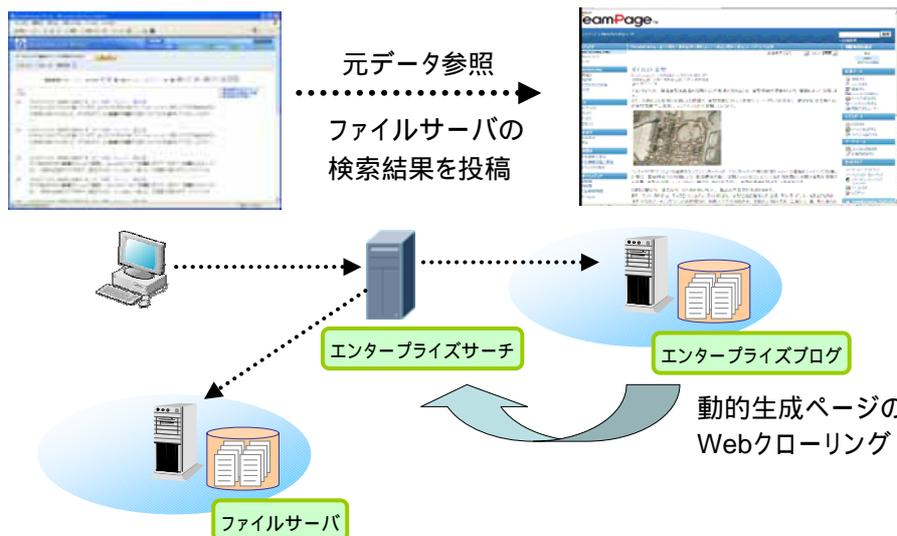
| 請求 | 請求日 | 請求先 | 請求件名 | 請求金額 |
|----|------|--------|-----------|-----------|
| | 8/30 | A銀行 | 支店LAN工事 | 3,800,000 |
| | 9/6 | B自動車 | 9月度保守費 | 8,250,000 |
| | 9/8 | 楽々電気工業 | コンサルティング費 | 1,350,000 |
| | 9/11 | 楽々商事 | 楽々電工向けPC | 450,000 |

| 受注 | 受注日 | 顧客名 | 受注件名 | 受注金額 |
|----|------|--------|-----------|------------|
| | 8/26 | C運輸 | 新システム機器一式 | 12,500,000 |
| | 9/4 | 楽々電気工業 | 9月度保守費 | 6,000,000 |
| | 9/5 | システム開発 | 楽々電工保守代行 | 2,500,000 |

- 項目に対し属性検索、類似検索を併用可
- 複数テーブルやViewをまたぐ検索も簡単

31

エンタープライズブログ連携



32

事例1: R&D系の膨大なデータを持つメーカーA社

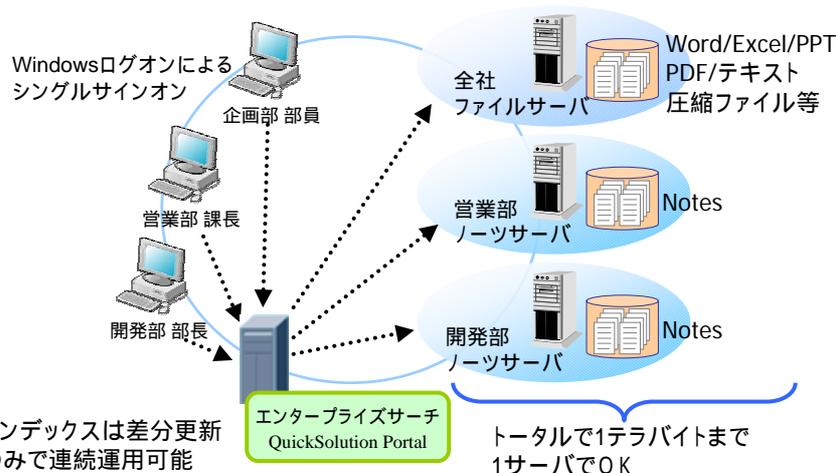
- 800GBのファイルサーバ
 - Word/Excel/PPT/PDF/テキストファイル/圧縮ファイル等
 - 部署ごとおよび開発プロジェクトごとにフォルダ作成
 - ADによるフォルダおよびファイルの権限を継承したい
- 200GBのノーツサーバ
 - ノーツDB、文書、ビューの権限を継承したい
- 連続稼動が必須
 - 定期的なインデックス再生成によるシステム停止は避けたい
- 検索サーバの台数を最小限にしてコストを抑えたい
- Windowsログオンによるシングルサインオン
- 利用者は数千人

35

事例1: R&D系の膨大なデータを持つメーカーA社

複数サーバの横断検索

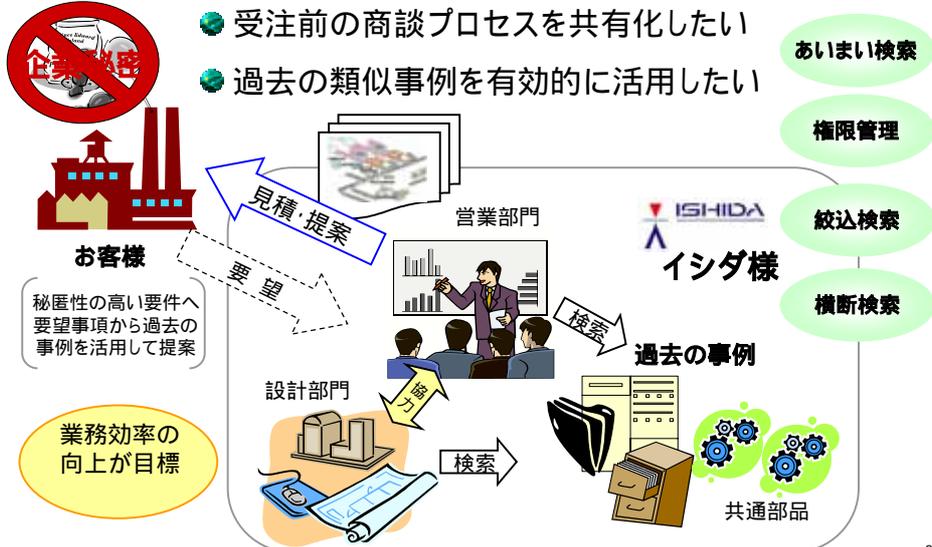
ADやNotesのアクセス権限継承



36

事例2：産機システム部門の課題

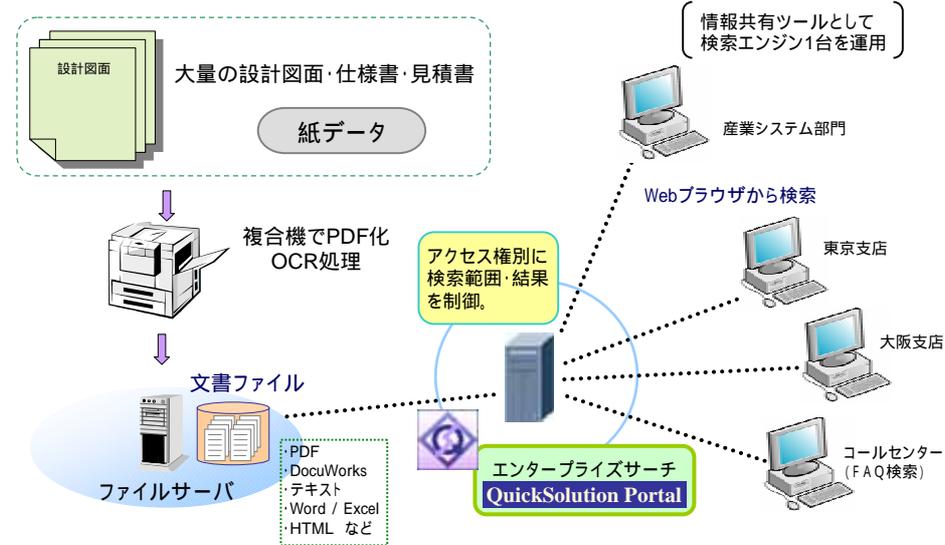
受注生産メーカーでの実践事例



37

事例2：システム構成図

受注生産メーカーでの実践事例



38

事例2: 導入のメリットと今後の課題

受注生産メーカーでの実践事例

- 完璧な認識精度ではないOCRにマッチ
(例)「!」、「1」、「ポテトチップス」、「ホテトチップス」
- 検索結果を見ながら、担当者の経験で絞込検索
(例)工番、機種、ユーザ名、計量精度などで絞込み
- インデックス生成中でも検索が可能
日中、文書をスキャン・追加しながらも検索可能
- コールセンターのFAQ検索にも活用

 業務効率の
大幅な向上
を達成

今後の課題

基幹システムとの連携

Net-It連携

検索結果のまとめ表示(要望)

画像メタデータの検索(要望)

 2008年2月
QSP 6.1で機能実装

39

事例3: RDBの高度な検索

- 特定フィールド重視
例:「タイトル」と「詳細説明」がある場合に
タイトルでヒットしたデータを上位に表示
- 同一フィールド複数検索種別
どこに何があるか分からない
場合の全文検索と
フィールドを指定した
属性検索を使い分け・併用
- 多重ソート機能
例:スコア アクセス数
新しい順
- カテゴリ集計
検索結果をカテゴリ等の
指定したフィールドの値で
件数カウント

 検索時のヒット件数が
数十万件でも高速処理
可能であること

データ件数:200万件



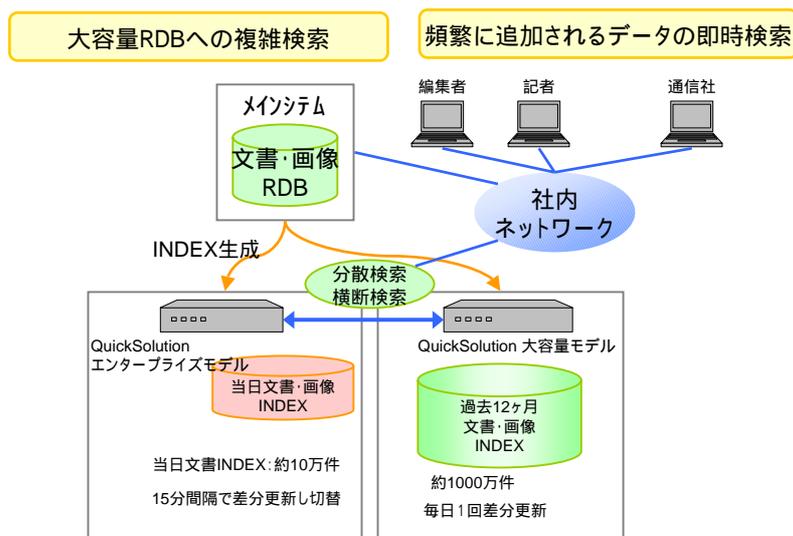
40

事例4: 膨大な記事を保有する報道B社

- RDB (Oracle) 1000万件を1秒以内に検索
- 表記揺れに対応した類似検索
+ カテゴリ、位置情報、日時等での範囲検索
- 更新データをできる限りリアルタイムに検索したい
- PCでの検索に加え、ケータイでの検索サービスも提供
 - 限られた画面サイズでのカテゴリ件数表示
 - 多彩なランキングおよびソート機能による表示順序の制御

41

事例4: 膨大な記事を保有する報道B社



42

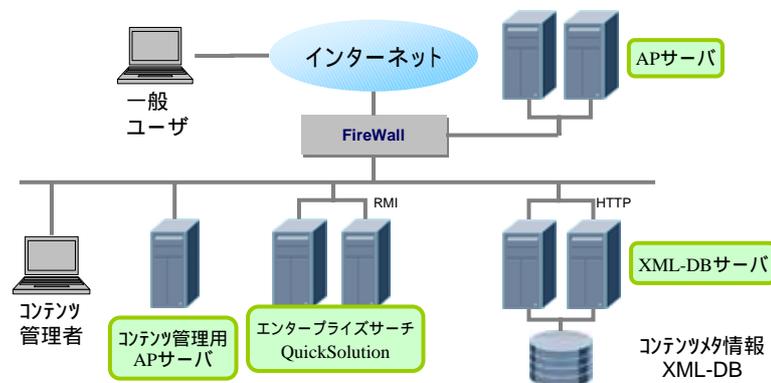
事例5：大規模コンテンツを保有する音楽系C社

- XML-DBに格納された数百GB～数TBの大規模コンテンツ検索
 - 新語、専門用語等の検索
 - 自然文検索
 - 表記揺れ対応
- 追加コンテンツのリアルタイム検索
- 用途・目的に合わせた検索画面の提供
 - 内部のコンテンツ管理者向けプロ向け検索画面
 - 外部の一般ユーザ向けの素人向け検索画面

43

事例5：大規模コンテンツを保有する音楽系C社

- ・類似検索および属性検索の両方を組み合わせて活用
- ・APIを用いてリアルタイム検索を実現
- 属性検索インデックス：リアルタイム更新
- 類似検索(全文検索)インデックス：「更新データを含めた検索機能」を使用
- ・一般ユーザ向け検索画面に加えて管理者向け専用検索画面を提供



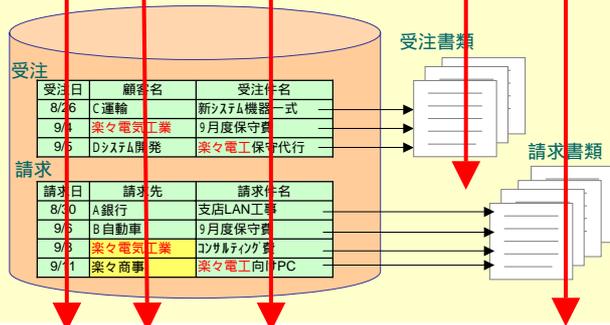
44

事例6：基幹システムと融合した検索イメージ

検索キーワード：

 9月の請求または受注で、
「楽々電気工業」関連


基幹システム



45

事例6：表記のゆれの吸収

- 検索したい文字列：「楽々MBA株式会社」
- 検索結果：0件

製品検索

検索キーワードを入力して下さい

検索

受注コード

受注先 楽々MBA株式会社

受注日

受注品名

検索

製品検索

検索キーワードを入力して下さい

検索

受注コード

受注先 楽々MBA株式会社

受注日

受注品名

検索



- ・スペルミス 「楽々MBA株式会社」
- ・大文字 / 小文字 「楽々MBA株式会社/楽々mba株式会社」
- ・全角半角 「楽々MBA株式会社/楽々MBA株式会社」
- ・文字のゆれ 「楽々MBA」 (**株式会社を省略**)

「楽々MBA株式会社」に
関する受注をとりだしたい
んだけど???

ユーザーが検索したいものがヒット!!!

46

事例6：全データ項目を一括検索

- RDBの場合、全項目を対象に一気通貫で類似検索できる
- 個別の項目に対するキーワード検索もできる

類似検索

キーワード検索

本文も検索対象

タイトルも検索対象

| No. | 質問番号 | 質問登録日 | 質問タイトル |
|-----|-------|------------|---------------------------------------|
| 1 | 1,002 | 2007-09-12 | FF2.0.0.4で日付型手動で初期値をセットする入力が出来ない |
| 2 | 1,007 | 2007-09-12 | PmsFormValueにSeiLinkで初期値をセットする入力が出来ない |
| 3 | 1,016 | 2007-09-12 | graphic型、vargraphic型の部分一致検索(OR検索)について |
| 4 | 1,021 | 2007-09-13 | WFとの差動連携時に日付項目の表示日付が切り替わってしまう場合があります。 |

47

事例6：検索結果の表示

- 類似検索とキーワード検索
- ファイル内検索 (text/csv/Excel/Word/PDF等主要形式)
- 最小スコアの指定
- スコアの表示、該当箇所の強調表示

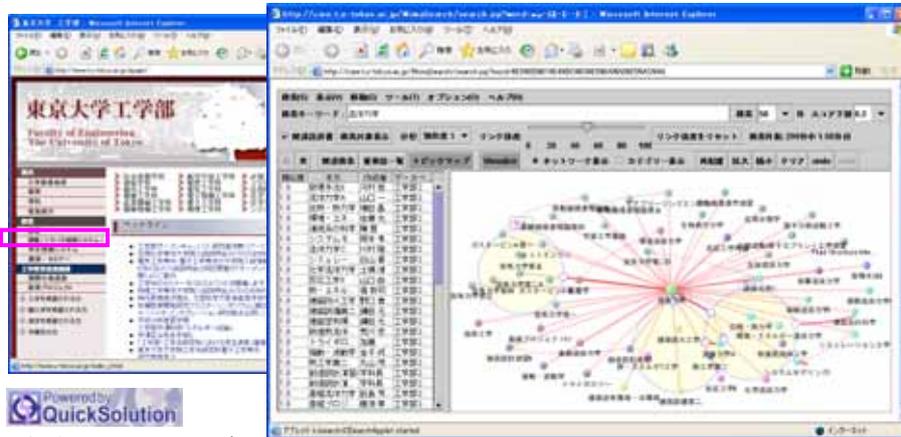
最小スコアの指定

該当箇所の強調表示

| No. | 質問番号 | 質問タイトル | スコア | テキスト |
|-----|-------|---|-------|-------------------------------|
| 1 | 1,021 | Matrixi (ターンでSeiLinkに値を渡す方法【問合せ内容】第三期完結の付録です。お世辞になります。バージョン番号 + Framework Ver. 1.41001 値を渡す方法はありません。【デジカメ】31041 | 1.000 | |
| 2 | 1,016 | graphic型、vargraphic型の部分一致検索(OR検索)について | 0.900 | OR検索の漢字子として認識するよに修正します。【デジカメ】 |

48

事例7：東京大学工学部「知の構造化」MIMA Search



可視化クラスタリング

- ・工学部やオープンコースウェアのシラバス検索などで採用
- ・最先端の言語処理により知識を階層化 類似化 可視化
- ・図示された講義の関連度を参考に、履修科目を容易に編成

【適用実績 / 応用例】

- ・特許マップ、論文検索
- ・発想支援、集中と抜けの発見
- ・企業内スキル/人材マップ等

49

お問合せ

住友電工情報システム株式会社

URL <http://www.sei-info.co.jp/>

E-mail qs-visitor@sei-info.co.jp

50