

リクルートにおける 検索エンジンSolrの活用

中野 猛

tf0054@r.recruit.co.jp

システム基盤推進室 / FIT / 株式会社リクルート

まだ、ここにはない、出会い。

アジェンダ

ご紹介

Solr?

利用状況

利用詳細

- 01) リクルートの紹介
- 02) Solrとは(概要と利用イメージ)
- 03) 利用サイト紹介(Solr利用サイト)
- 04) 採用した理由(ドリルダウンでSolr!)
- 05) 採用までの道のり
- 06) 利用の詳細(インフラ)
- 07) 利用の詳細(アプリ/開発容易化の施策)
- 08) 今後の展望(大規模/v1.3)

リクルートのご紹介

リクルートの会社概要

[ご紹介](#)[Solr?](#)[利用状況](#)[利用詳細](#)

- 創業 昭和35年3月31日
- 資本金 30億264万
- 売上高 4,436億円(2006年3月期)
- 営業利益 1,297億円(2006年3月期)
- 従業員数 6,298名
- 平均年令 30.8歳



リクルートの事業概要

ご紹介

Solr?

利用状況

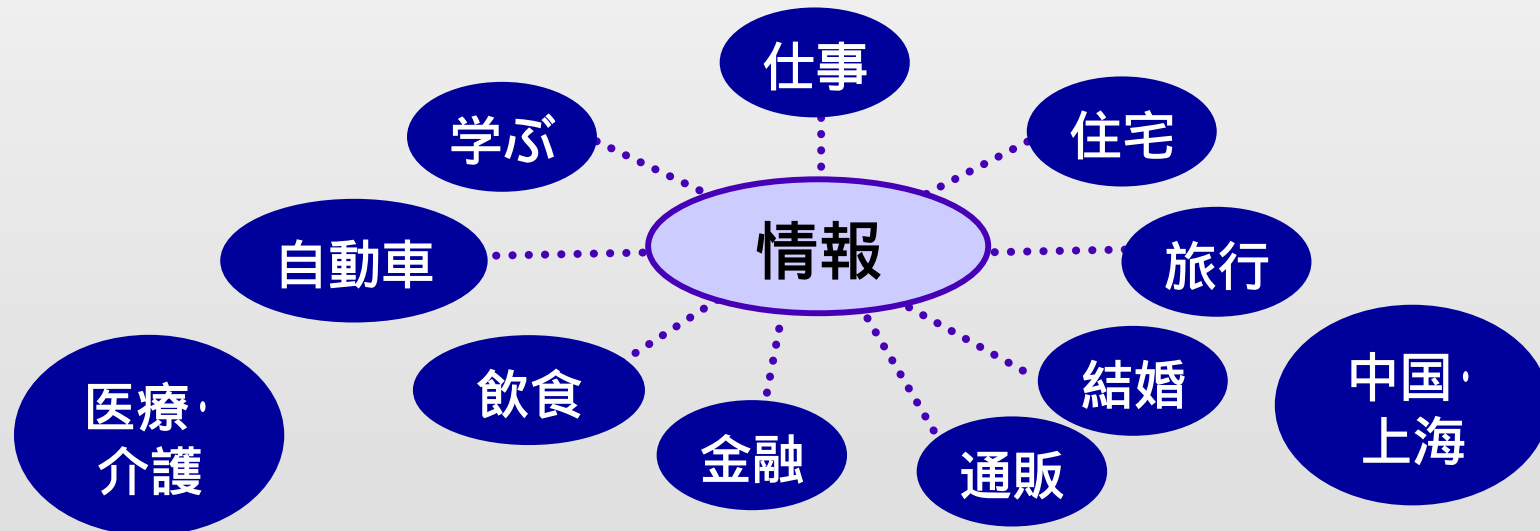
利用詳細

- 人材総合サービス(HR)

- 求人情報サービス、人材斡旋サービス、個人のキャリア形成支援サービス、人材派遣etc.

- 商品とカスタマーとマッチングサービス(販促系)

- 学び / 住宅 / 旅行 / ブライダル / 出産育児 / 自動車におけるマッチングサービスetc.



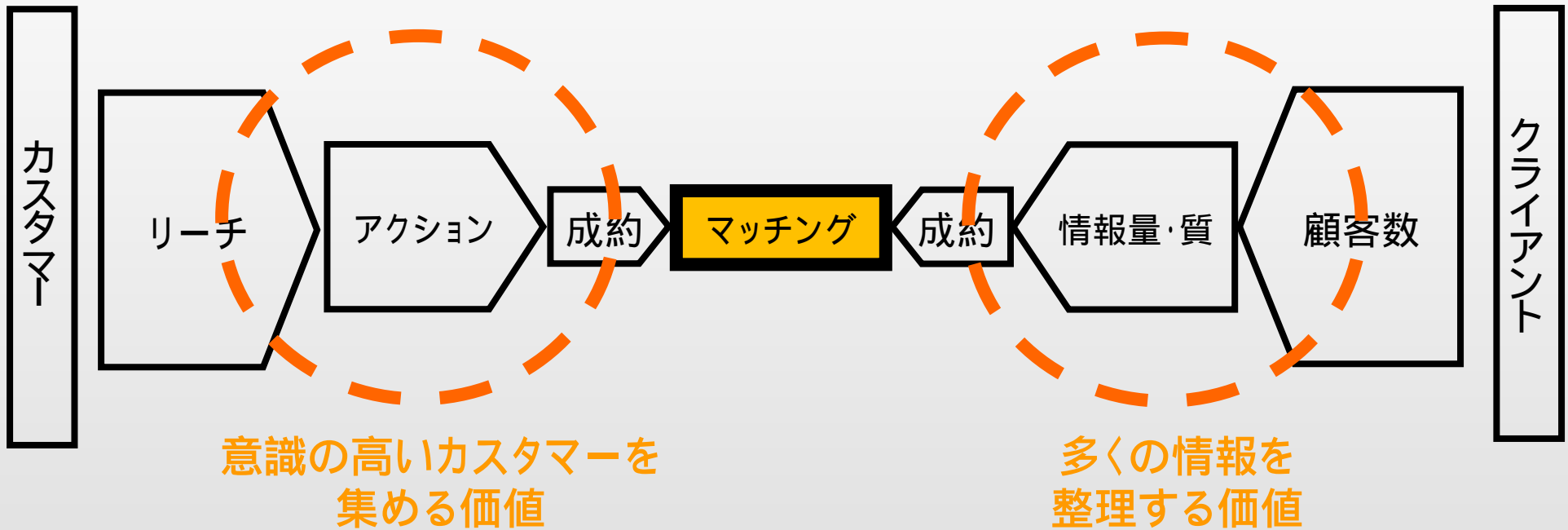
リクルートのビジネスモデル

ご紹介

Solr?

利用状況

利用詳細



FITとは

ご紹介

Solr?

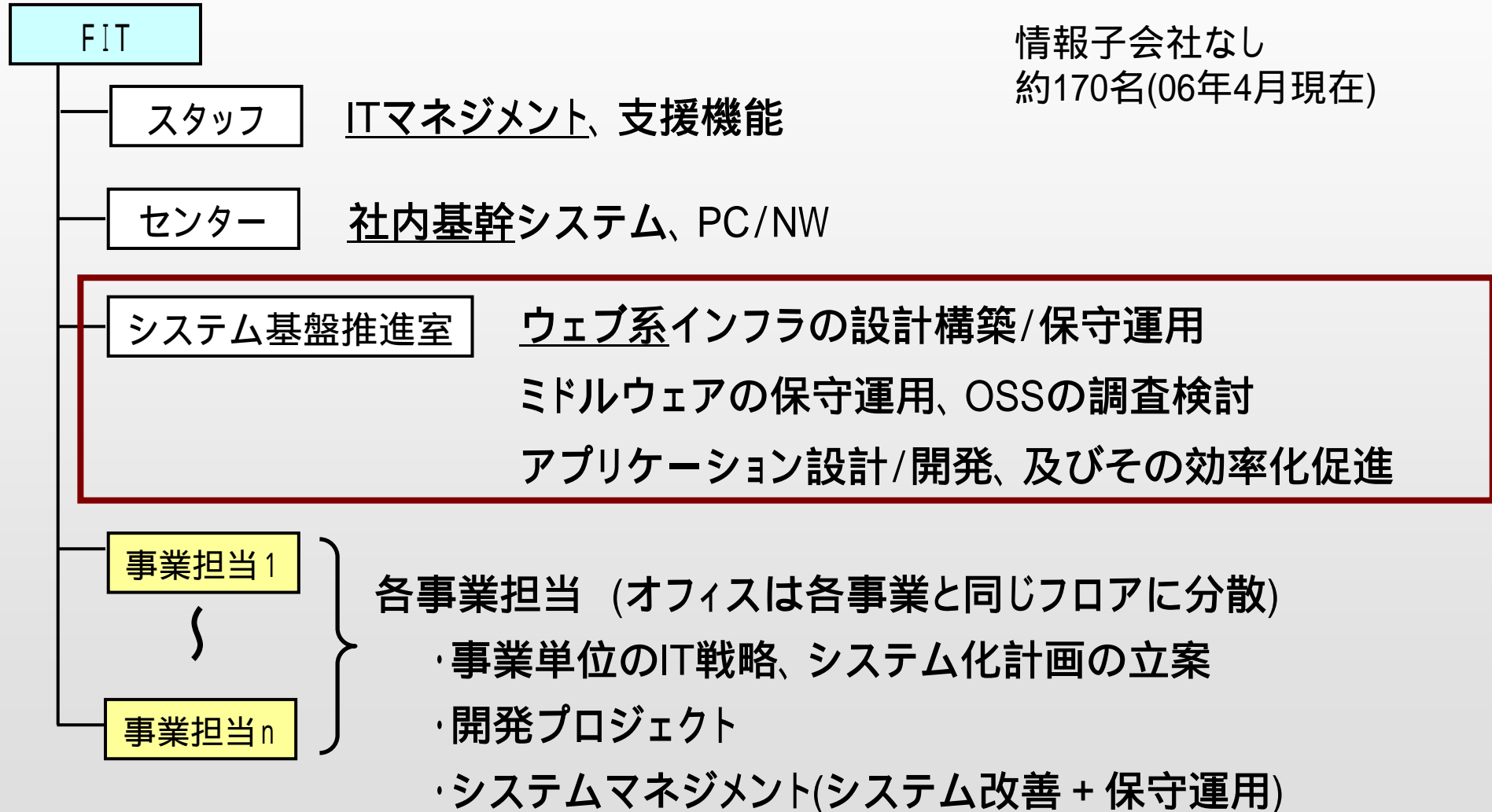
利用状況

利用詳細

● Federation of IT : 全社横断の情報システム部門

情報子会社なし

約170名(06年4月現在)



自己紹介

ご紹介

Solr?

利用状況

利用詳細

- 中野猛(ナカノタケシ)

- 2000年入社

- これまでの仕事内容

- 新ISIZEインフラ設計/構築/保守運用
- 某新規事業の開発&保守運用
- R25ウェブサイトの開発&保守運用

- OSS含む新技術の調査 & 社内導入推進

- PostgreSQL, Solr, ZABBIX, Introscope, etc.

- 次期ウェブサイト設計/開発基盤の検討

- 画面/アプリ/インフラを通して全体で最適化!



検索エンジンSolr

Solrとは

ご紹介

Solr?

利用状況

利用詳細

- OSSの検索エンジンミドルウェア
 - サーバアプリケーション
 - CNET社内製造後オープンソースに
 - apache.orgのTOPドメインプロジェクト
- Solrの良いところ
 - **ファセット機能** (ダイナミックドリルダウン!)
 - 結果並び順など挙動の解明可能 (OSS故)
 - 必要なHWが小規模で安価 (検索機能に特化)
- Solrの弱いところ
 - ロボット等でのクローリング取込み
 - 検索語の集計や分析



Solrとは

ご紹介

Solr?

利用状況

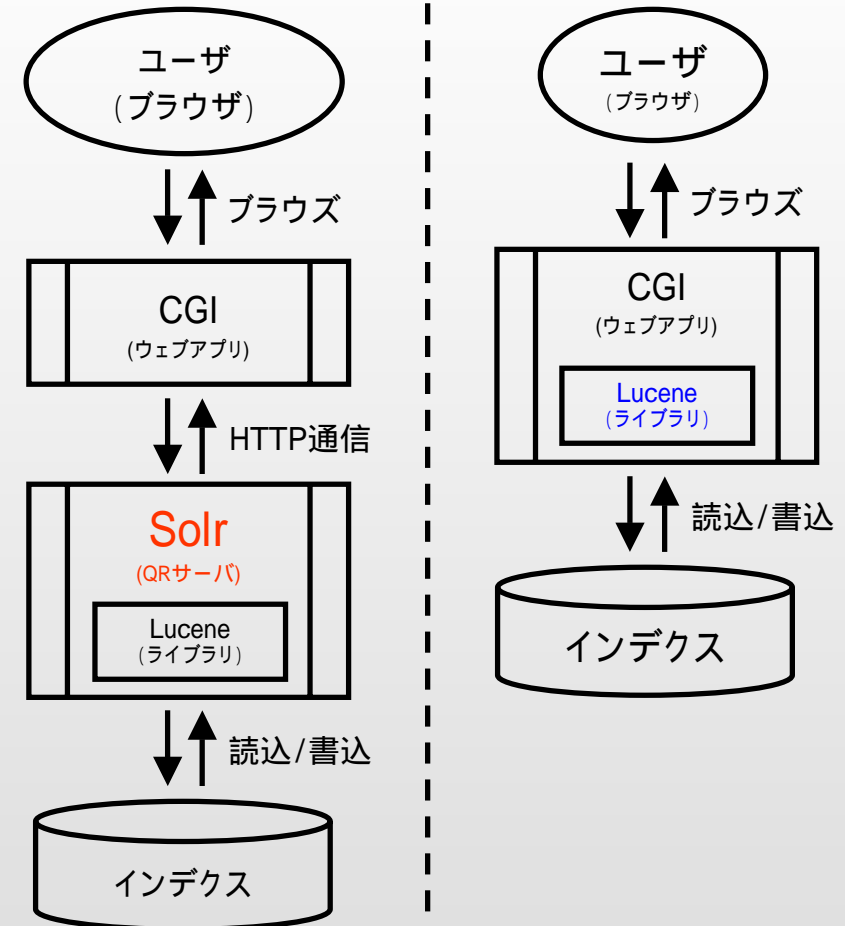
利用詳細

●形態

- Tomcatにデプロイ
- XML@HTTPでI/F
- コアライブラリはLucene

●サーバアプリ

- 開発工数の削減
- 維持運用者の機能分離
- 障害切り分けに有利



Solrとその他検索エンジン

ご紹介

Solr?

利用状況

利用詳細

- オープンソースの検索エンジンは数多存在
 - 独自にI/O機能を有する(DB活用型/サーバ型)
 - ライブラリでCGI等プログラムに組み込む(組み込み型)

➤ DB活用型

- Tritonn
- Luida
- など

➤ サーバ型

- Solr
- JiroSearch
- など

➤ 組み込み型

- Hyper Estraier
- Namazu
- Lucene
- など

● Solrのその他ポイント

- DiggやInternetArciveでの利用実績
- 拡張可能なAPIが多く組み込まれた構造
- ファセット機能を考慮したキャッシュを持つ

利用サイトのご紹介

Solrを活用したサイト例

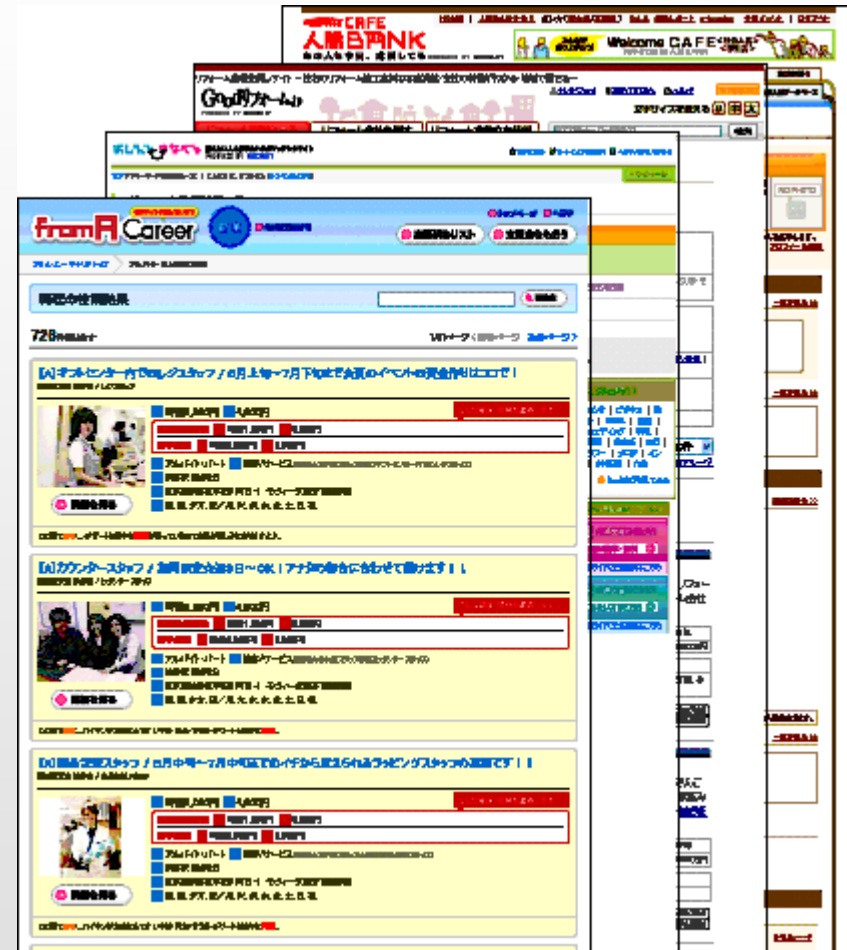
ご紹介

Solr?

利用状況

利用詳細

- 人脈バンク
 - グッドリフォーム
 - おしえるまなべる
 - FromAキャリア
- +
- 2サイト開発中...



Solrを選んだ理由

- 弊社で必要とされているのは
 - 日本語キーワードでの検索機能よりも主に欲しい機能は**ファセット**
 - 商用ミドルの載せ換えも考えるとサーバ型が有利

- ファセットと
 - ダイナミック
 - 情報を先

The screenshot shows the WiseNut search engine interface. At the top, the logo 'WiseNut Search Exactly' is visible. Below the search bar, the results are categorized under 'WiseGuide categories for "diving"'. The categories are listed in two columns, each with a plus sign icon and a search link. The number of results for each category is shown in parentheses. The category 'Diving Cruise' has 5 results and is circled in red. Below the categories, the first search result is displayed, starting with '1. Scuba Diving Board - Online Dive Community - Scuba Equipment, Travel a...'. The URL for this result is 'http://www.scubaboard.com/'.

Category	Count	Search Link
Scuba Diving	107	[search this]
Learn to dive	3	[search this]
BALI DIVING	9	[search this]
Key West	3	[search this]
Diving Snorkeling	7	[search this]
Scuba Gear	9	[search this]
Diving Cruise	5	[search this]
Red sea	3	[search this]
Shark Diving	7	[search this]
Diving holidays	12	[search this]
Diving Thailand	11	[search this]
Cave Diving	3	[search this]
Diving The worlds	3	[search this]
Bahamas Diving	3	[search this]

1. [Scuba Diving Board - Online Dive Community - Scuba Equipment, Travel a...](http://www.scubaboard.com/)
 ... Join ScubaBoard and over 65,000 divers discussing Scuba **Diving** topics
 Read dive articles where you can learn about the latest news, deep & wr
<http://www.scubaboard.com/>

Solrを使う前は...

- ファセットの種類
 - 特定カラムをユニークに集計
 - 特定カラムを自由に集計("1000円台の商品"など))
- これをDBでやると大変...
 - 前者はgroup-by **リソース負荷が高い**
 - 件数は毎回条件付でcountする **もしくは集計テーブルを自前でメンテ**
- Solrではインデクス時にある程度**事前計算**
 - 単語一覧 > インデクスに一覧表を持つ
 - 件数カウント > インデクスにそのまま入っている

利用開始までの道のり

これまでの流れ

[ご紹介](#)[Solr?](#)[利用状況](#)[利用詳細](#)

● ステージ

- 2006年度4Q 調査/検証を開始
- 2007年度1Q 準備を開始
- 2007年度2Q 案件へ適用
{
- 2008年度1Q 大規模対応を検討

これまでの流れ

[ご紹介](#)[Solr?](#)[利用状況](#)[利用詳細](#)

- **ステージ**

- 2006年度4Q **調査/検証を開始**

- 2007年度1Q **準備を開始**

- 2007年度2Q **案件へ適用**

）

- 2008年度1Q **大規模対応を検討**

これまでの流れ

ご紹介

Solr?

利用状況

利用詳細

● ステージ

- 2006年度4Q 調査/検証を開始
 - 複数の商用検索ミドルウェアが稼動
 - ノウハウや人的リソースが分散
 - 包括契約ではなくサポートコストがうなぎ上り
 - 社内各所でLuceneを調査(書籍が出た為)
 - ノウハウや人的リソースを集約
 - 社外リソースの開拓と一本化
- 2007年度1Q 準備を開始
- 2007年度2Q 案件へ適用
}
- 2008年度1Q 大規模対応を検討



調査/検証

ご紹介

Solr?

利用状況

利用詳細

● 日本語での利用

● 変更無くとも動きはする

- Solrチームの牧野勝氏の記事参照(Software Design 2007/12)

● 文章の単語分割が必要

- 形態素(MeCabのJNIバイディングを利用)

我輩は猫である = 我輩 は 猫 で ある

- N-gram(Solrの標準ライブラリで対応可)

我輩は猫である = 我輩 輩は は猫 猫で であ ある

● どちらが良いも言えない

- 前者は精度は高いがモレも在る(後者はその逆)

調査/検証

ご紹介

Solr?

利用状況

利用詳細

● 調査

● 構造解析

- SolrのクラスやLuceneインデクスの構造

● 負荷試験

- 800万件程度までテストし商用エンジンと比較

● その他

● 運用監視

- 監視すべき項目を検討(キャッシュヒット率他)
 - 管理画面をパースして監視アプリへ流すことに
- 運用が必要な項目を想定
 - 単語追加や順位等の問い合わせ(後者は想定ほど無し)

利用の促進

ご紹介

Solr?

利用状況

利用詳細

- 商品掲載サイト
 - じゃらんに代表されるサイト
 - 要求SLは極めて高い
- 集客に注力するサイト
 - 広告掲載などに代表されるサイト
 - 要求SLの許容幅が比較的大きい(インプレッション保障故)
- 後者から始めて徐々に展開
 - 結果的にはさっそく商品掲載サイトへ
 - (Solrに起因した)サイト停止は無し
 - (今期)更なる大規模サイトへ展開予定

利用の詳細

Solrの利用推進体制(社内)

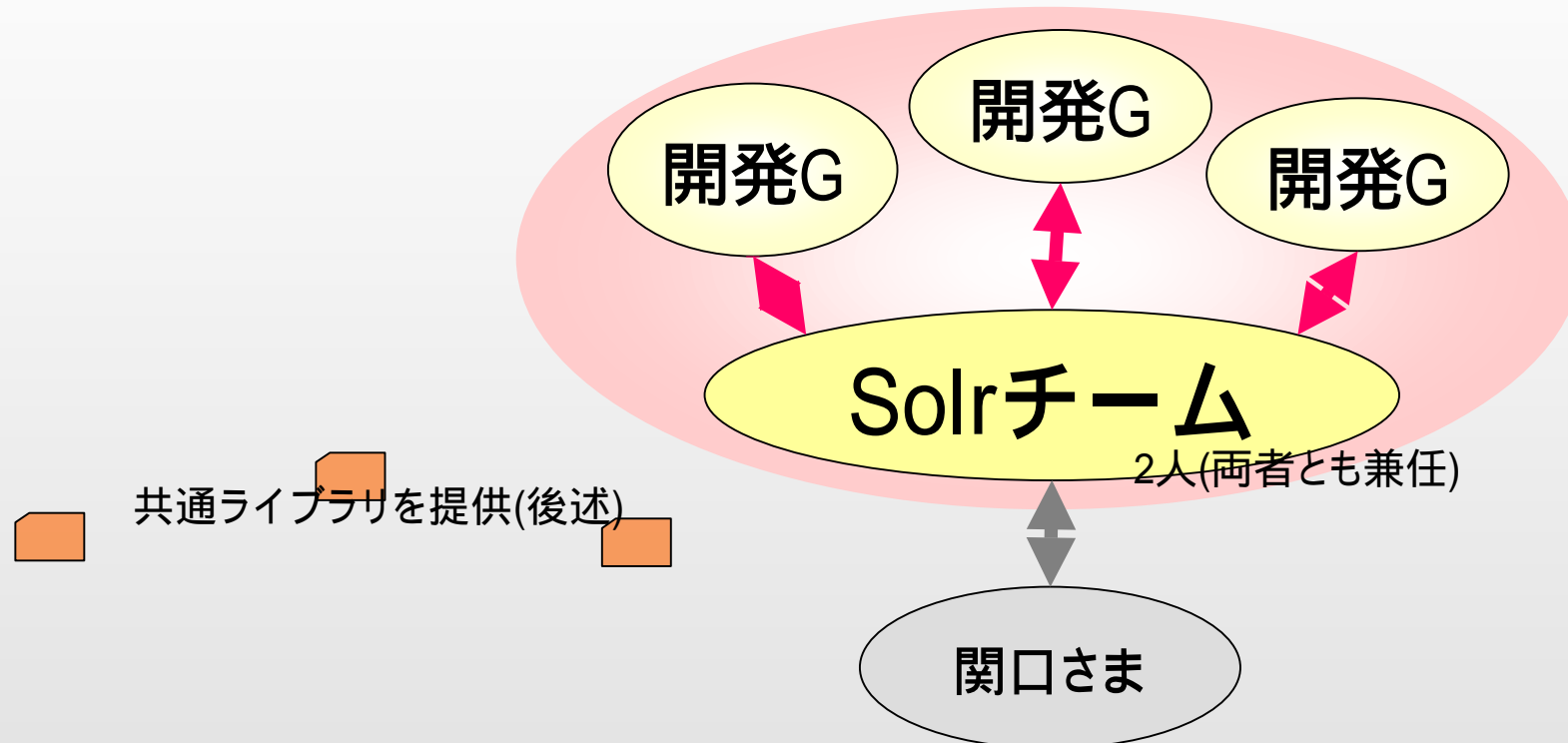
ご紹介

Solr?

利用状況

利用詳細

- OSSに強いメンバを調査時からアサイン



- この体制のまま開発サポートチームへシフト
- これにより開発利用がスムーズに立ち上げ可能に

Solrの利用推進体制(社外)

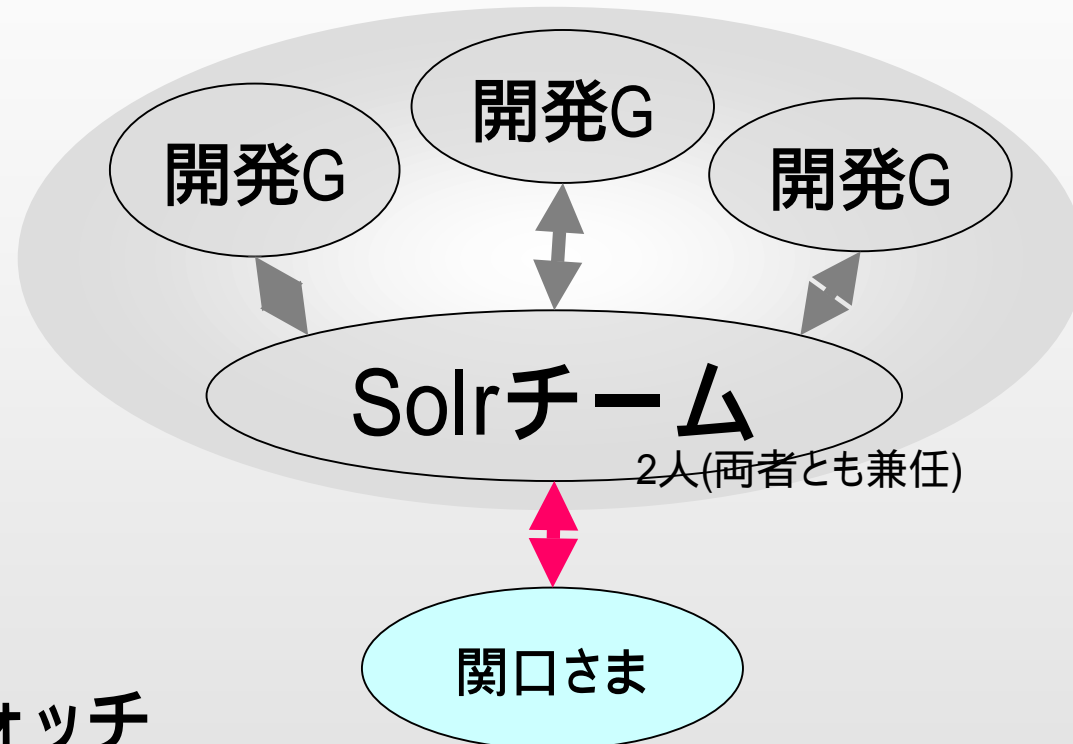
ご紹介

Solr?

利用状況

利用詳細

- Solrの構造解析をメインに調査



- コミュニティのウォッチ
 - 新規公開パッチの機能紹介(1個/週程度)
- 不明挙動のQA/調査/パッチ作成
 - 実績では月1件あるかないか程度

アプリの開発サポート(スキーマ)

ご紹介

Solr?

利用状況

利用詳細

- 1インスタンスは1スキーマで固定
 - DBで考えると1つのTBL構造に固定される
 - Solrには設定XMLファイル内に記載して渡す

#	カラム名	タイプ	単語分解	データ格納
0	ID	Int	-	×
1	NAME	String	MeCab	
2	PRICE	SortableInt	-	×
3	CATEGORY1	String	2-gram	×
4	CATEGORY2	String	-	×
5

- 元データが複数ある場合は汎用的に作る必要有
- 各カラムによって単語分解方法を変更可能
 - 商品名など検索モレが許されないところはN-gram等

アプリの開発サポート(ワークシート)

[ご紹介](#)[Solr?](#)[利用状況](#)[利用詳細](#)

- スキーマ設計にはワークシートを準備
 - 開発者のスキルはまちまち
 - スキーマはサイト開発者が作成
 - Solrチームがレビュー(ワークシートで会話)
 - エクセルにから直接設定(XML)を生成
 - 設定ファイル作成の手間を軽減

アプリの開発サポート(ライブラリ)

[ご紹介](#)[Solr?](#)[利用状況](#)[利用詳細](#)

- 共通ライブラリを準備(Java)
 - 文書の投入や検索が可能
 - HTTPで通信していることを意識させない
 - 2つのSolrを使い分け可能(冗長化)
 - コミュニティ標準(?)のSolrJをカスタマイズ
 - エラー処理などを社内ライブラリ向けに
 - 結果をオブジェクトにマッピング
 - 開発者に優しい

インフラは相乗り設備

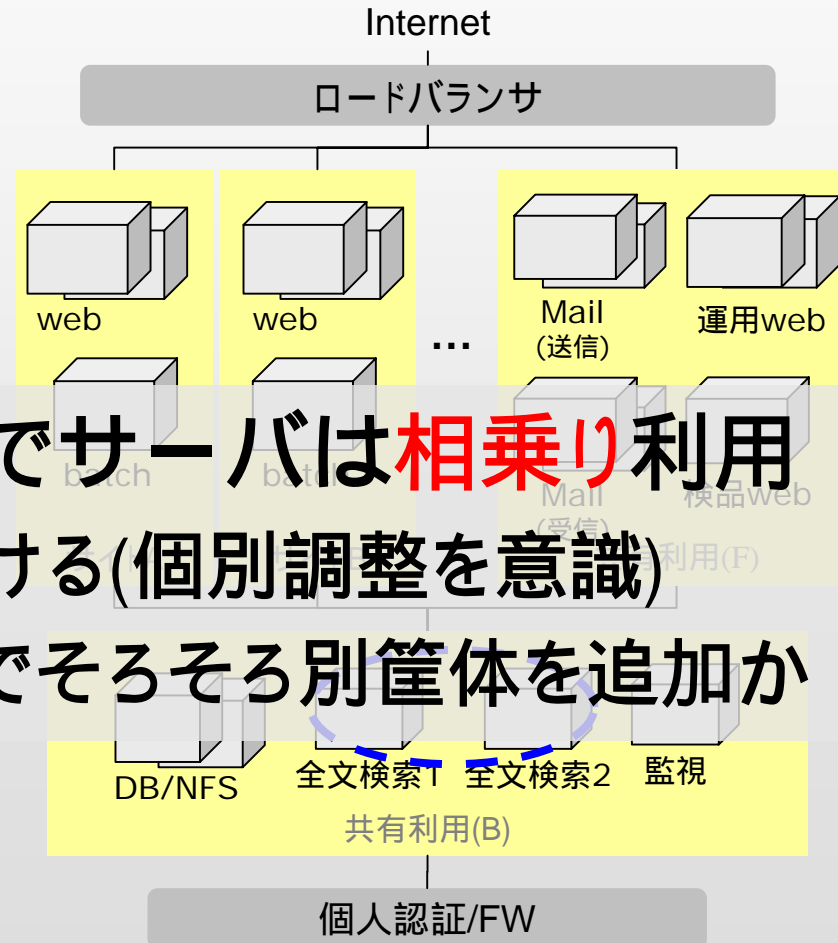
ご紹介

Solr?

利用状況

利用詳細

- Solrには個別にサーバ筐体を割当て
 - 「2-Xeon+4G+RAID1」程度



- 複数のサイトでサーバは**相乗り**利用
 - Tomcatで分ける(個別調整を意識)
 - 現在4サイトでそろそろ別筐体を追加か

インフラでの監視と運用

ご紹介

Solr?

利用状況

利用詳細

● 監視

- Solrは管理画面をパースして流し込み(前述)
 - 最近JMXで取れるよう機能追加がなされた
- 監視/モニタツールZABBIXにデータを集約
 - オープンソースのツール
 - DBデータ監視等と同じ環境で統合モニタリング

● 運用

- 通常のJVM/Tomcatアプリと同
 - 初期導入や簡単な設定調整
 - セキュリティパッチの適用



今後の展開

リクルート内部から

ご紹介

Solr?

利用状況

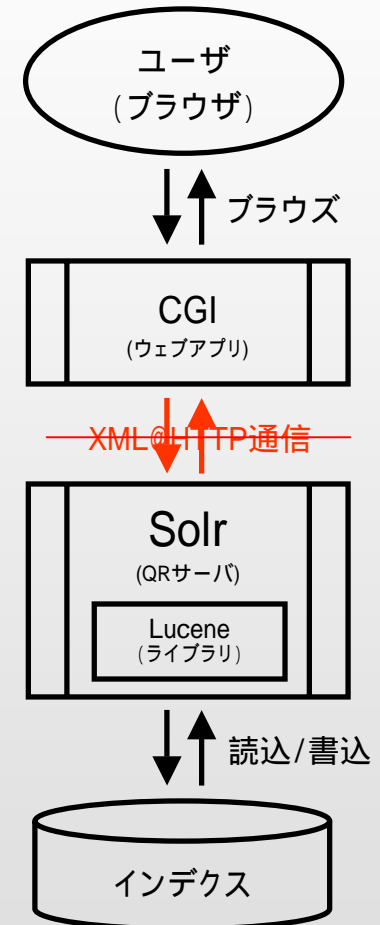
利用詳細

● 利用の拡大

- 高アクセス/大規模サイト
 - Solrの機能的キャパより人材
- 既存サイトへの検索アドオン
 - 自サイトのクローリング要件もあり
 - アプリ構造がシンプルに

● 新版ライブラリのリリース

- 問い合わせ数が増え高負荷に
 - XMLパースは大変 > バイナリへ変更
 - プラグイン的に結果返却機能を差換え可
 - ハイライトは大変 > 検討中(n-gram?)
 - MeCabでJNIも調整次第か



Solrとして

ご紹介

Solr?

利用状況

利用詳細

- 新版(v1.3)がもうすぐ
 - 分散ノードでの串刺し検索
 - 文書数では現行(v1.2)でも**1億強** = 1インスタンス
- DBと連携
 - ThinkITに関口さんの記事(在庫が無い>順位を下げる等)
 - DB連携型の開発容易性を取り込む
- 評判検索
 - 企画ありきで世に問う
 - 技術ドリブンでのトライ