Solrスゲェ！

株式会社 EC ナビ　システム本部
EC ナビラボグループ
春山　征吾
Seigo_Haruyama@ecnavi.co.jp

EC ナビ・EC ナビラボの紹介

- EC ナビ　(http://ecnavi.jp)

- EC ナビラボ (http://labs.ecnavi.jp/)
  - 検索 (search) と情報共有 (share) をキーワードを軸に，次世代のソフトウェア技術，インターネットサービスについての研究開発を行なっています .

今回提供する API の紹介

- Buzzurl API
- http://labs.ecnavi.jp/developer/buzzurl/api/
- Buzzurl ブックマークの投稿・検索ができます .

- NavicSearch API
- http://labs.ecnavi.jp/developer/navicsearch/api/

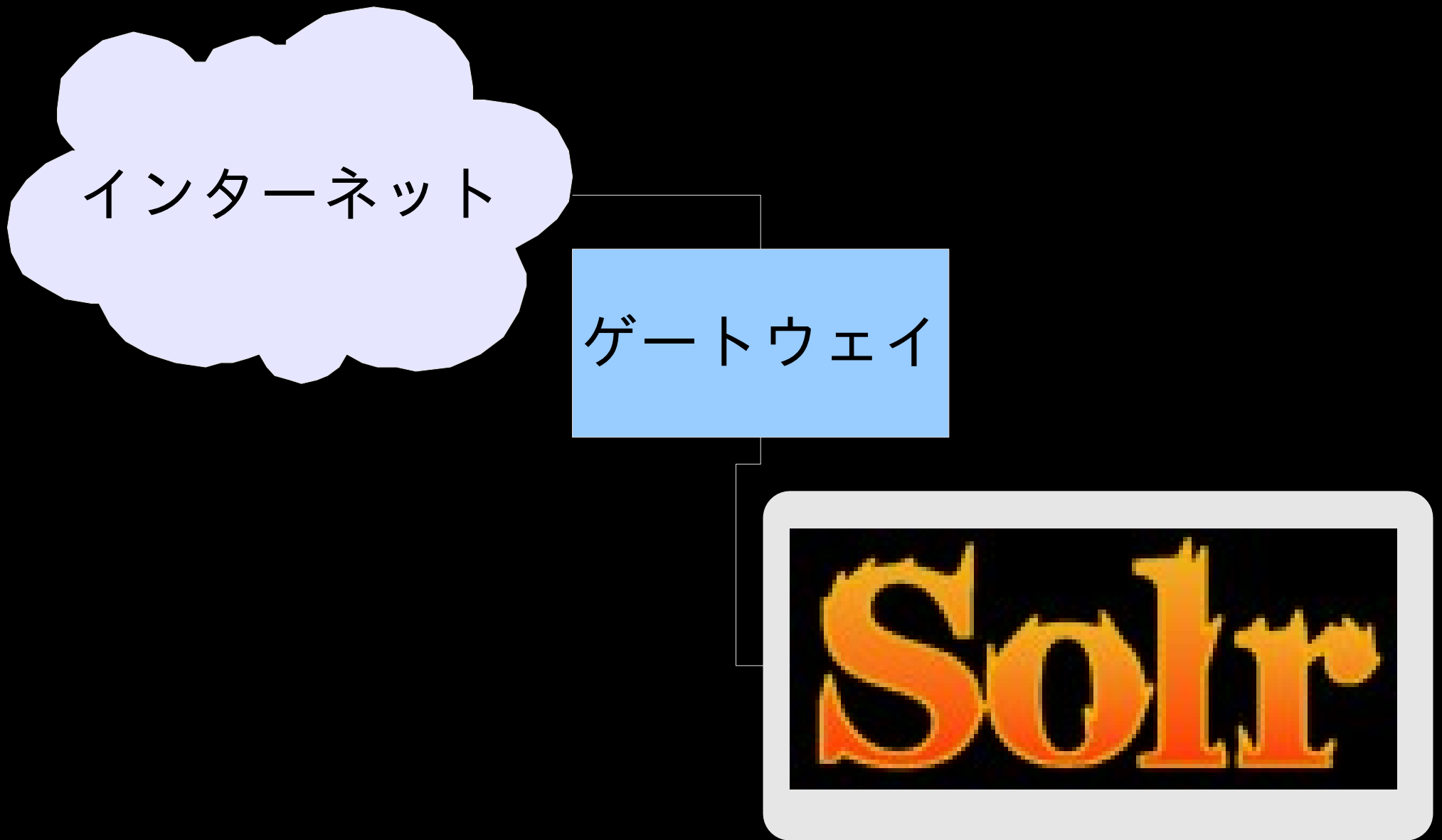
NavicSearch API の特徴

- EC ナビで取り扱っている 1500 万件以上の商品の検索ができます .
- 検索クエリの例 :
  - 「 wii 」で検索
    - http://api.ecnavi.jp/shopping/navic_search/select?q=wii
  - 「自転車　折りたたみ」で検索
    - q=自転車　折りたたみ
- デモサイト！
  - http://s-tanno.net/

NavicSearch API (α) の注意事項

- 現在開発中の EC ナビの商品検索機能のプレビュー版です.
  - 商品データは実際のものですが，データの更新は不定期です .
  - プレビュー版であるためほとんどの場合， EC ナビで検索した検索結果と異なる結果を返します.
- 予告してもしくは予告せずにサービスを一時停止する可能性があります .

NavicSearch API のシステム構成



Solr とは？

- 「全文検索エンジンライブラリ Lucene をベースに、管理画面やキャッシュ機構を取り入れたアプリケーション」(Wikipedia より)
- 発音 : Solar と同じ .
- 利用例 : Digg やインターネットアーカイブ .
- Java で作られています .
- 2008 年 5 月に Lucene の開発者の一人の関口宏司さんが Solr の開発者になりました .
  - 今後日本における利用も期待されています .
  - Buzzurl でも Solr を使い始めました !

全文検索システムにも

- Tritonn (MySQL + Senna)
- Ludia (PostgreSQL + Senna)
- Namazu
- Hyper Estraier

...

とありますが

なぜ Solr を使うことにしたかというと

それは Solr がスゲェから

- ここからひたすら Solr をたたえます .

高速

- 特にチューニングしない Solr と特にチューニングしていない Tritonn にデータをつっこんでみた (100 万件).
  - Solr が 10 倍以上速かった .
- 検索の速度については十分なデータは取っていませんが , Buzzurl の状況を見ると Tritonn の同等以上のようです .

スケールアウトが容易

- 同じインデックスを複数のマシンで持ってサービスできる (レプリケーション)
- 異なるデータセットを持つ (パーティショニングされた) 複数の Solr を一括検索できる

- 理論上いくらでもスケールアウトできます .

機能拡張が容易

- 結構きれいに設計されているので，機能拡張が容易です．
- 後で示すように，実際に機能を追加して使っています．

アプリケーション作成も容易

- HTTP (GET or POST) で検索式入力
- 出力フォーマットを指定可能
  - XML, JSON(P)
  - Ruby, Python, PHP

なので様々な言語から簡単に扱えます .

「10分で簡単！RailsとSolrの全文検索デモ構築」
(関口さんの会社　RONDHUIT の資料)

ファセット (Facet) 機能

- ファセット (Facet) とは
  - カテゴリやキーワード，価格，日付といったものを用いた分類
- Solr は簡単にファセットごとの件数を出力してくれます .
  - NavicSearch API でもカテゴリごとに商品を検索することが可能
  - 例 :
    「デンドロビウム」で検索してカテゴリ名でファセ

最近の Solr の動き

- 現在の Solr のリリースバージョンは　1.2
- EC ナビで利用している Solr は 1.3 nightly (1.3 は 7 月末リリース予定)
- 開発が活発でばんばん機能追加されてっています .
- Nightly 版でないと日本語などを扱える CJKTokenizer が付いてません .
  - 1.2 での NGramTokenizer はあるのでまったく日本語が扱えないというわけではありません .
  - Tokenizer とは，検索のためのインデックスを作るために文書を Token に分割するもののことです .

EC ナビ独自の Tokenizer (といってもパクリだけど)

- Solr の CJKTokenizer は (すくなくとも 6/18 時点のでは)
  - 空のトークンが末尾に入る
  - いわゆる半角カナを無視する (これは仕様)

という問題があるので，独自に作りました .

- といっても，http://twistbendcoupling.net/501/cjktokenizer で Public Domain で公開されていたものをパクりました .
- その名も　ECNaviTokenizer.
  - 半角カナの濁点・半濁点も適当にします .

まとめ

Solr スゲェ！

みんなも使おう

Solr