Building a thesaurus for product search in ecommerce

From Magento Elastic Suite to PCU

04/07/2017, RISE CAEN

PCU

Overview

- **Why product search ?**
- **Now - Magento Elastic Suite primer**
- **How MES Thesaurus answers ecommerce-specific needs**
- **Tomorrow – PCU : Machine Learning for ecommerce**
- **Questions**

The speaker

- **Marc Dutoo, R&D projects lead at Smile, the leading EU Open Source service provider**
- **PCU project coordinator, Data / API / Cloud expert**

PCU

Why product search ?

Why product search ?

- In ecommerce, search is very important

  - Typically **60%** of product buys come from it, and only 40% from category / shelves browsing...
  - but their management cost doesn't reflect that !
- A specific, very concrete criteria of search results being right or wrong : whether the customer **buys** its products.
- This financiary incentive curbs everything:

  - Autocompletion is paramount
  - Boosting search fields
  - Scoring searched products
  - Rescoring results
  - Enriching results
  - ...thesaurus building

PCU

Never empty results

- Never empty results
  - In other search domains, for instance knowledge management or entreprise search, an empty result is a good answer, meaning that the knowledge or document is not yet there and adding it would be an improvement
  - But in ecommerce, there should never be 0 result, because the customer must **never be in a dead end**
- => “push” generic, “hot” products to the customer : most viewed, most bought, discounted, newest, available, hot brands...
  - branches out to recommendation algorithms (the other axis of ecommerce)
- But also curb search algorithms to have as wide returns as possible
  - correction, query expansion i.e. **thesaurus**

Products differ widely across domains

- A lot of **different features** across categories
  - Books : only author, editor
  - Smartphone : size, camera resolution, memory, network, color...
  - Overall, 80 distinct filters are used in 800 mo page view logs
  - Product categories also can be searched
  - And always : price, discount, but also brand, availability, store, description...
- Whose contribution to search varies. Ex. description is often **too wide** :
  - Book description : can cover as many topics as book do
  - T-shirt description : “this blue t-shirt goes very well with yellow trousers”
- The solution
  - Search in all fields, but allow filtering those specific fields => combined search + filters functionality
  - And allow to configure **separately** how each field / feature contributes to search results : boost title, don't look in book description...

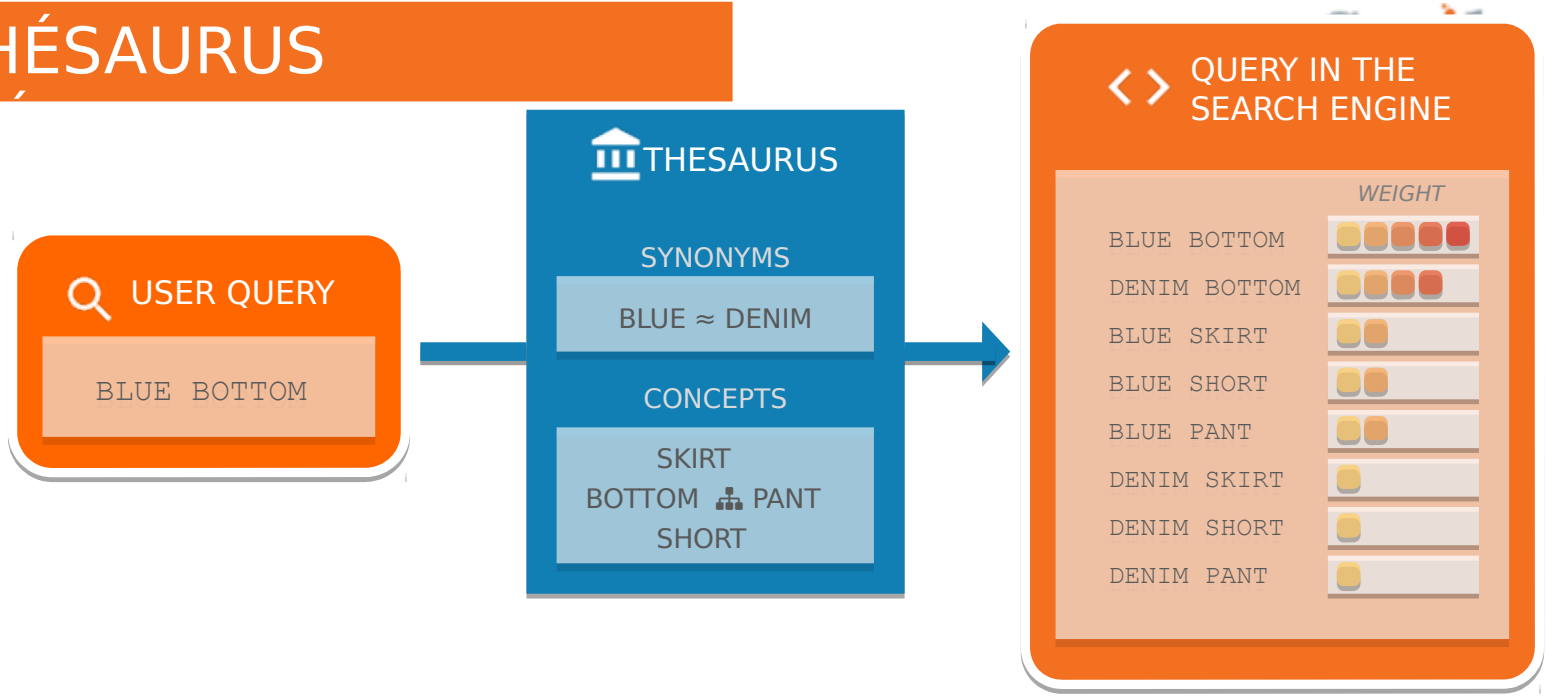
PCU

Now – Magento Elastic Suite primer

Now – Magento Elastic Suite primer

- **Magento** : leading Open Source ecommerce platform
- **ElasticSearch** : Open Source distributed search platform and ecosystem, easy to set up and integrate in business applications, built on Apache Lucene
- **Magento Elastic Suite** (MES) : Open Source “searchandising” solution by Smile, the leading Open Source solution provider in Europe
  - Searchandising : mixes search and product selling optimisation techniques and marketing
- Features :
  - Search (field boost, thesaurus...)
  - Facettes (i.e. filters)
  - Merchandising (product placement...)

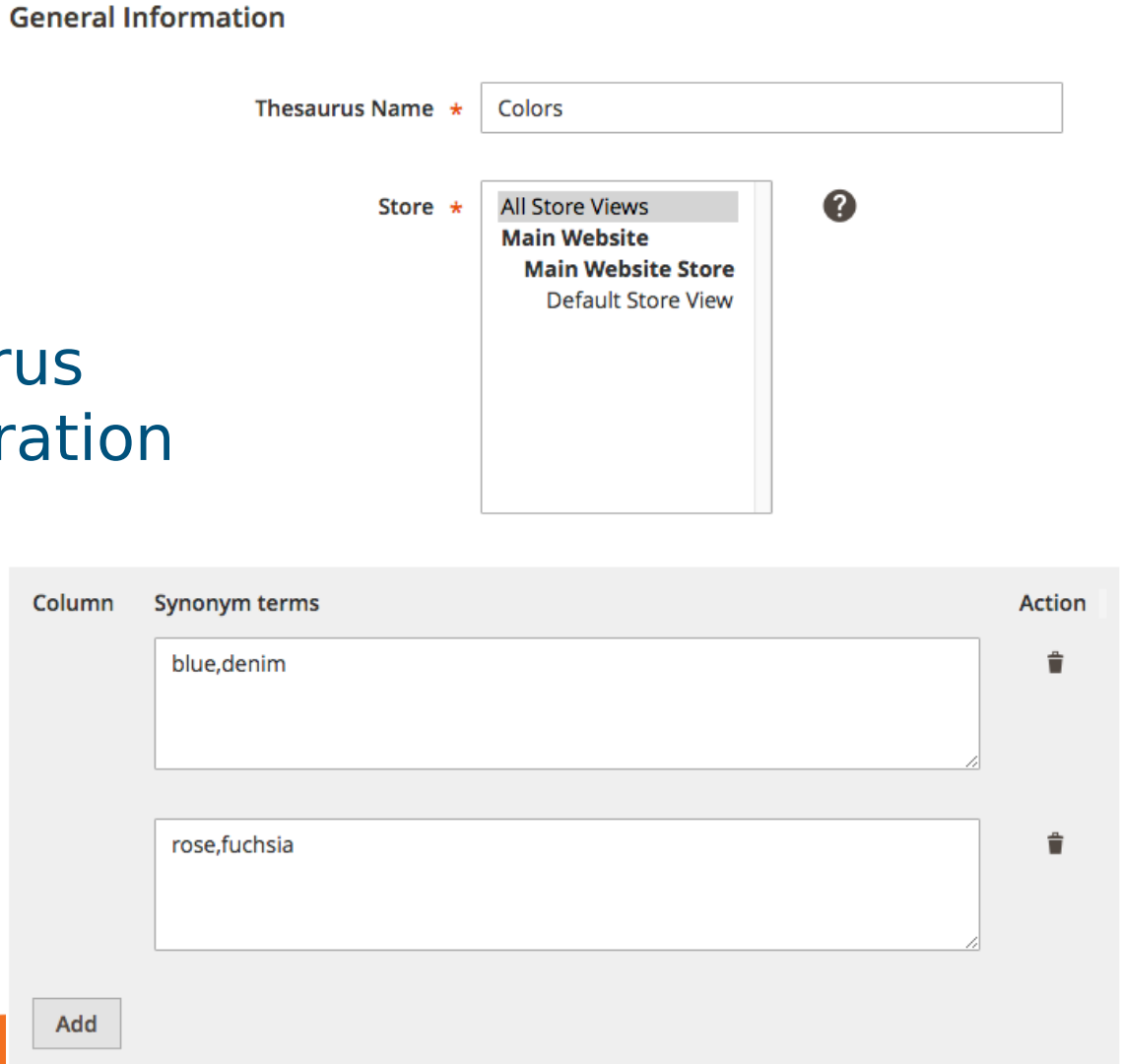
MES THÉSAURUS



Des règles sémantiques pour optimiser le moteur de recherche

- **Définissez des synonymes et des concepts adaptés à votre catalogue produit**
- **Le moteur de recherche utilise ce thésaurus pour créer des requêtes voisines de la requête de l'utilisateur**
- **Tous les résultats sont consolidés en utilisant un poids adapté**

Thesaurus configuration



MES Thesaurus, for ecommerce specific needs

PCU

MES Thesaurus – practical examples

- **color** : denim, sky blue, cyan... : thesaurus of blue
- **clothes** : jeans <= trousers
- **DIY** / bricolage : grass => seeds
- HOWEVER iphone - galaxy : incorrect example
  - of a ”high end smartphone” thesaurus rule, because iphone is a different world than Android and Apple fans are very loyal to the brand !
  - => rather a “high end smartphone” category
- ... i.e. very **specific** to each kind of product, catalog, vendor

MES Thesaurus – building process

- When the ecommerce website is being built, MES experts talk with the vendor to help him configure it
  - Products, categories, thesaurus...
- 3 weeks after it has gone “live”, MES experts study how it is used and patch configuration
  - By looking in search analytics (Magento backoffice or Google Analytics) :
  - In **searches with empty or few results**, find most entered terms : do they need synonyms to be added to the thesaurus ?
  - Also seldom clicked categories...

Ecommerce search beyond thesaurus

- Search of a **lot of words** : several searches are done and combined
  - Of all words together first – but higher risk of returning no result (also because higher risk of spelling mistake)
  - Then of each single word, then of all pairs of 2 words, then 3
  - Combined by giving less weight to those last ones

Tomorrow : Machine Learning for ecommerce with PCU

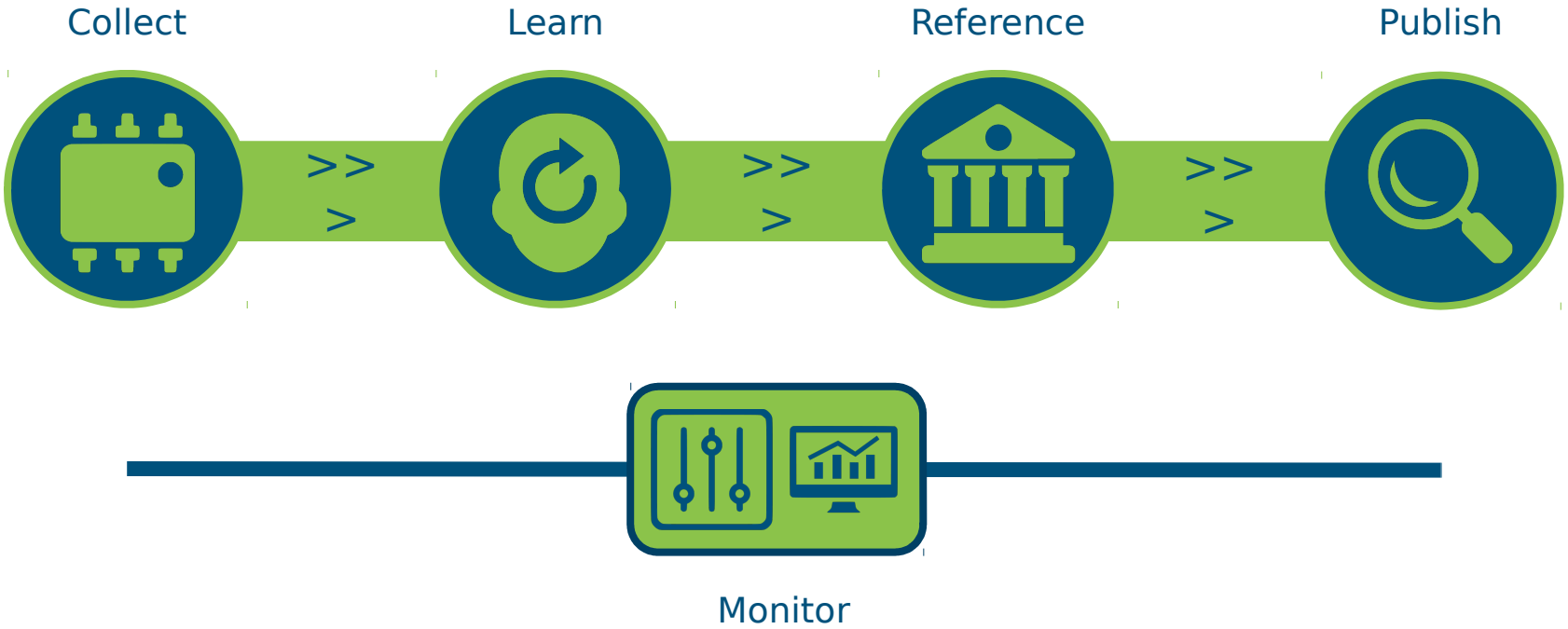
PCU

Factsheet - Unified Knowledge Platform

- 6 partners over 2017-2019, sponsored by the French ministry of Industry
- In order to democratize Big Data, so that every company will be able to add value to its own core business thanks to its existing data :
  - a Big Data / Machine Learning / semantic module to enrich any business application
- As showcased in 2 use cases :
  - E-commerce (up to digital in store) & B2B
  - Enterprise search
- Thanks to:
  - A factory of Machine Learning and Semantics-enriched search engines
  - state-of-the-art and new algorithms analyzing user behaviour
  - end-to-end event-driven data processing workflow
  - an open source, best-of-breed, unified, flexible and extensible approach

PCU

Overall architecture



PCU

Partners and stakeholders

Smile : coordinator, architecture, ecommerce

Paris 13 : Machine Learning, semantics

ESILV : pipeline, semantics

Proxem : text & opinion mining, B2B

Wallix : enterprise search experience

Armadillo : integration & mgmt API & UI

Financial sponsors : BPI, IDF

Cluster : System@tic

Key Machine Learning outputs - ecommerce

- Ecommerce : **recommendation** algorithms that learn from user behaviour, for advertising but also search autocompletion
  - Most searched product filters / features
  - Most searched query terms, predict their evolution (for autocompletion)
  - Collaborative filtering on product views
  - More widely, most viewed or sold products or features and their evolution...
- Ecommerce & B2B : also
  - Detect buy intent
  - predict sales
  - predict churn
  - Suggest or train customer segments, using classification

Key Machine Learning outputs - search

- **Named Entity Recognition** (topics & aspects)
  - For supervized enrichment of ontology
  - up to opinion mining, for polarity of opinion about aspect (good or bad)
- Allowing to transform fulltext search into **structured** search :
  - “samsung tv” => type:tv and brand:samsung
- Thesaurus that **learns** from user behaviour :
  - if a lot of people search “jeans”, but don't click on any result and rather search “trousers”, and only then click on a product => “jeans to trousers” query expansion should be added to thesaurus
  - Using search query co-occurrence
- **Entreprise search**, beyond files & classification :
  - Search competences and experiences...

Outputs

- **Generic platform**
  - Unified, flexible, extensible, best-of-breed-based, API-managed
  - Along with a set of standard connectors, data pipeline elements and Machine Learning (ML) and text mining algorithms
- **Use cases and products**
  - E-commerce (product, deployed at Smile early adopter customers), B2B (deployed at Smile & Proxem)
  - Enterprise search (product, deployed at each partner's)
- **Open Source Ecosystem**
  - Ties with integrated technical components' communities as well as derived business-specific products
  - Home of platform examples, tryout and adoption

https://pcu-consortium.github.io/
http://magento-elastic-suite.io - http://www.smile.fr
Contact : marc.dutoo@smile.fr

Questions ?

Thank you for your attention !