Článek:
Charron, B., Hirate, Y., Purcell, D., Rezek, M. (2016). Extracting semantic information for e-Commerce. *The Semantic Web – ISWC 2016, 2, 273-290*. Doi: 10.1007/978-3-319-46547-0 27

Důvod výběru:
Praktické využití ontologie, respektive taxonomie v IT (e-commerce)

Autoři:
Bruno Charron
Yu Hirate (https://www.researchgate.net/profile/Yu_Hirate)
David Purcell
Martin Rezk

Abstract:
Rakuten Ichiba uses a taxonomy to organize the items it sells. Currently, the taxonomy classes that are relevant in terms of profit generation and difficulty of exploration are being manually extended with data properties deemed helpful to create pages that improve the user search experience and ultimately the conversion rate. In this paper we present a scalable approach that aims to automate this process, automatically selecting the relevant and semantically homogenous subtrees in the taxonomy, extracting from semi-structured text in items descriptions a core set of properties and a popular subset of their ranges, then extending the covered range using relational similarities in free text. Additionally, our process automatically tags the items with the new semantic information and exposes them as RDF triples. We present a set of experiments showing the effectiveness of our approach in this business context.*
(Charron, Hirate, Purcell, & Rezek, 2016, p.273)

Rakuten Ichiba E-shop/marketplace

Japanese electronic commerce and Internet company based in Tokyo, Japan. Its B2B2C e-commerce platform Rakuten Ichiba is the largest e-commerce site in Japan and among the world's largest by sales. Revenue 2015 ¥713.5 billion > příjem 165 miliard / 2015 (wiki)

e-commerce,
eBooks, travel,
banking,
securities,
credit card,
e-money,
portal and media,
online marketing,
professional sports,
(Charron, Hirate, Purcell, & Rezek, 2016, p.275)

200 000 000 produktů
Taxonomie 40 000 tříd
2001 (OWL je 2004)
Japanese texts
Terabajty dat

Představme si tedy velikou Alzu.

BUS case
Taxonomie 40 000 tříd > potřeby rozšiřovat (doménová oblast, keywords při vyhledávání, chování zákazníků) > manuálně rozšiřováno (nutná doménová znalost, chyby, mnoho času, drané) > automatizace.

Nikoli všepostihující, ale UX

For instance, the class Wine has among its subclasses Red Wine, White Wine, and Wine Accessories. Clearly the properties for Wine and Wine Accessories are very different. Thus, trying to extract property-values from all the items in the Wine category directly leads to unintuitive (although correct) results, such as "yellow" for the property "color". (Charron, Hirate, Purcell, & Rezek, 2016, p.276)

Postup – jak problém řeší
1) Extrakce chování uživatele na webu (logy, dolování z prohlížeče)
2) Výběr struktur (stromů)
    a. Velký BUS dopad
    b. UI > objev nových produktů (potenciální nové nabídky)
    c. Homogenní pro uživatele
3) Výběr relevantních vlastností
4) Extrakce triples

Relevantní stromy
Extrakce stěžejních vlastností
Extrakce reprezentativních hodnot
Pro každou věc dohledání vlastností
Univerzální řešení

Layers:

Vstupy: ontologie, logy, popis předmětů
Computing: Cluster | Hive & Spark na HDFS
Procesing: P-V extrakce, Model training, MeCab Tokenizer, Gensim Word2Vec | Subtrees selector, Keywor Ranker

Extension: Semantic gate, Triple generator, Cleaner-aggreagtor-bootstraper

Výsledky (benchmarky) v článku

Tímto přístupem nahrazují klasický recommendation engine (nebo ho obohacují) a staví rovnou celou stránku včetně doporučených nabídek, příbuzných produktů a relevantních vlastností produtků pomocí ML => Pomocí ML generované sránky e-shopu.