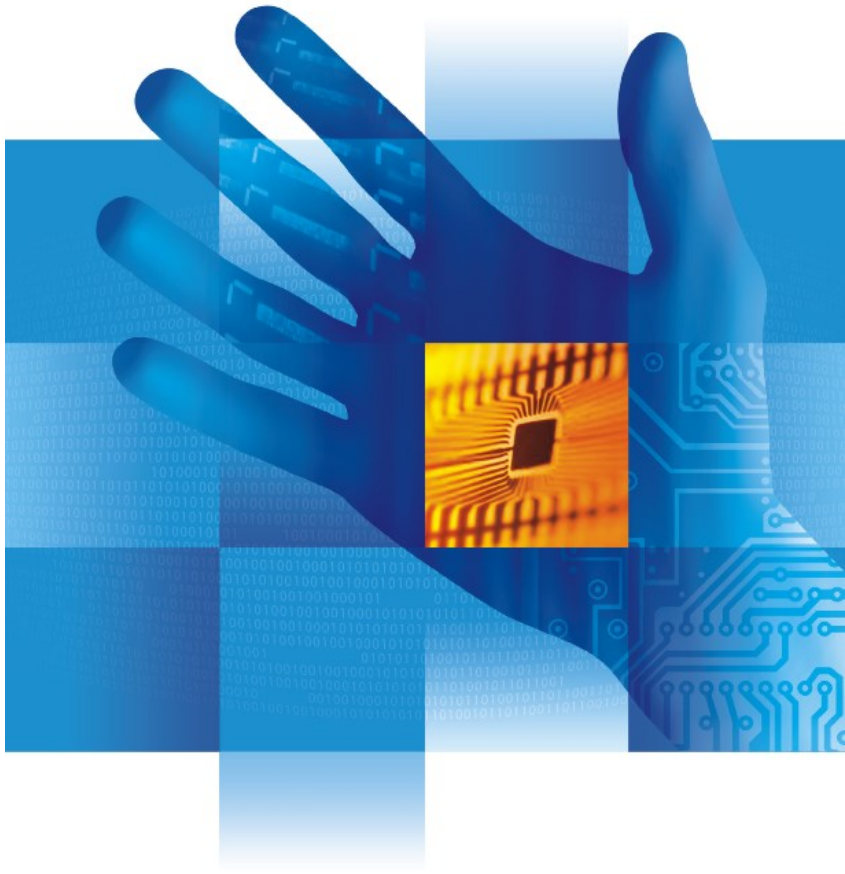
Dolování dat pro personalizaci webu

Petr Šaloun
VŠB-Technická univerzita Ostrava
FEI, katedra informatiky



Automatická personalizace a dolování (vytěžování) dat (1)

- Schopnost personalizovaného systému upravit obsah a doporučené položky vede k tomu, že musí být schopen určit, co uživatel potřebuje.
- To je založeno na předchozí nebo současné interakci s uživatelem.
- Personalizační úkol může být viděn jako problém predikce: systém se musí pokusit předpovědět uživatelovu úroveň zájmu ve specifických obsahových kategoriích, stránkách, položkách a seřadit je podle předpovězených hodnot.

Automatická personalizace a dolování (vytěžování) dat (2)

- Dále je úkol doručování personalizovaného obsahu často vyjadřován v pojmech odporučovacích úloh, kde systém doporučuje položky s nejvyšší odhadovanou hodnotou zájmu.
- Obecně může být personalizační systém viděn jako mapování uživatelů a položek na množinu zájmových hodnot.

Přístupy k personalizaci

Z pohledu architektury a algoritmů se personalizační systémy dělí do tří základních skupin:

- systémy založené na pravidlech
- systémy filtrování obsahu
- spolupracující filtrovací systémy

Personalizační systémy založené na pravidlech

Rule-Based Personalization Systems

- pracují s manuálně nebo automaticky vytvořenými rozhodovacími pravidly, které jsou používány pro doporučování položek uživatelům
- časté využívání systémů s manuálními vytvářením pravidel
- umožňují administrátorům specifikovat pravidla, často založená na personálních charakteristikách uživatelů

Filtrovací systémy založené na obsahu

Content-Based Filtering Systems

- uživatelský profil reprezentuje obsahový popis položek, ve kterých uživatel dříve vyjádřil zájem
- obsahové popisy položek jsou reprezentovány sadou vlastností nebo atributů, které charakterizují položku
- uživateli jsou doporučeny položky, které jsou odpovídající svou podobností uživatelskému profilu

Spolupracující filtrovací systémy

Collaborative Filtering Systems

- Zahrnují porovnávání hodnocení objektů (filmů, produktů atd.) aktuálním uživatelem s hodnocením podobných uživatelů (nejbližších sousedů), aby vytvořily doporučení pro objekty, které dosud aktuální uživatel neviděl, nebo nehodnotil.
- Tradiční používanou metodou je klasifikační přístup *k* nejbližších sousedů (*k*NN – *k*-Nearest-Neighbor), který porovnává profil cílového uživatele s historickými profily jiných uživatelů, aby našel prvních *k*, kteří mají stejné zájmy.

Přístupy k profilování uživatelů

- Všechny přístupy k personalizaci založené na dolování dat vyžadují kolekci dat, která přesně odpovídá zájmům uživatelů a jejich interakcí s aplikacemi a položkami.
- Systémy založené na pravidlech a filtrovací systémy založené na obsahu budují individuální model uživatelových zájmů a používají tento profil k úpravě budoucí interakce pouze s tímto uživatelem.

Data mining přístup k personalizaci

- Web usage mining není specifický algoritmus – standardní data mining cyklus
- Vytěžování používání webu: *automatické zjišťování a analýza vzorů v proudu kliků a souvisejících údajů získaných nebo vytvořených v důsledku interakce uživatele s webovými zdroji na jedné nebo více webových stránkách.*
- Cílem vytěžování využití webu je zachytit, modelovat a analyzovat vzory chování a profily uživatelů při práci s webem.
  Zjištěné vzory jsou obvykle reprezentovány jako kolekce stránek, objektů, nebo zdrojů, které jsou často navštěvovány skupinami uživatelů se společnými potřebami a zájmy.

Cíle vytěžování používání webu

- lepší porozumění návštěvníkům webu
- automatická adaptace a zpřístupnění personalizované funkčnosti

Personalizace webu má tři fáze:

- příprava a transformace dat
- zjišťování vzorů
- doporučování

Jen doporučování probíhá v reálném čase.

Reprezentace uživatele a zájmu

uživatel $u \in U$

$U = \{u_1, u_2, \cdots, u_m\}$ máme m uživatelů

položka $I = \{i_1, i_2, \cdots, i_n\}$ n položek

profil uživatele

$$u^{(n)} = \langle (i_1, s_u(i_1)), (i_2, s_u(i_2)), \cdots, (i_n, s_u(i_n)) \rangle,$$

s_u (nrozměrný vektor uspořádaných dvojic)

kde funkce

přiřadí uživateli u skóre zájmu o položku

(zájem může být nulový)

Předzpracování dat

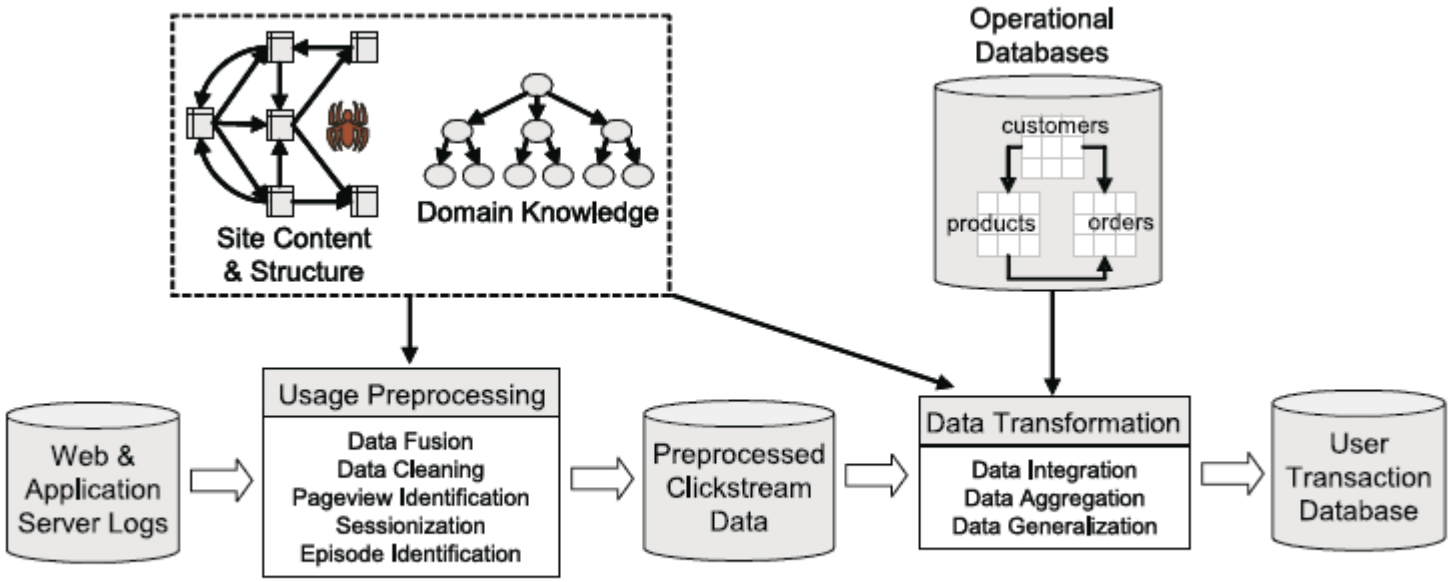


Fig. 3.1. Summary of the primary tasks and elements in usage data preprocessing.

Závěr

- Klíčovou částí personalizačního procesu je generování uživatelských modelů.
- Nejběžnější používané uživatelské modely reprezentují uživatele jako vektor hodnocení nebo používají sadu klíčových slov.

Závěr

- I když jsou k dispozici multidimenzionální nebo ontologické informace, data jsou obecně mapována na samotnou tabulku, která více odpovídá většině technik dolování dat a strojového učení.