

Porovnávanie inštancií ontologických konceptov pre potreby personalizácie

Anton Andrejko and Mária Bieliková

Ústav informatiky a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita, Ilkovičova 3, 842 16 Bratislava
{andrejko,bielik}@fiit.stuba.sk

Abstrakt Porovnaním a ďalším skúmaním spoločných a odlišných atribútov konceptov, môžeme získať zaujímavé informácie o používateľovi. V príspevku opisujeme prístupy k určeniu podobnosti konceptov s cieľom zohľadniť vo výsledku jedinečnosť používateľa. Podľa definovaných prahových hodnôt rozdeľujeme atribúty konceptov na množinu pozitívnych a negatívnych atribútov. Z týchto atribútov a z explicitného ohodnotenia, napr. záujmu používateľa o koncepty, môžeme usudzovať o charakteristikách používateľa.

1 Prečo porovnávať

Informácie dôležité pre potreby personalizácie môžeme získať viacerými spôsobmi, napr. sledovaním činnosti používateľa, analýzou záznamov webových serverov, vyžiadanim si informácie priamo od používateľa a pod. [1] K informáciám o používateľovi sa môžeme dopracovať aj analýzou obsahu, ktorý sa prezentuje používateľovi [4].

Iniciatíva webu so sémantikou so sebou prináša reprezentáciu ontológiami. Základ pre ontológie tvoria koncepty [7]. Koncept chápeme ako množinu s jednoduchými vlastnosťami (nazývanými aj atribúty), ktoré sú naviazané na zodpovedajúce koncepty. Ich inštancie zodpovedajú objektom reálneho sveta. Existujú dva pohľady na koncept [6]: *extencionálny* (angl. extensional) a *intencionálny* (angl. intensional). Pri extencionálnom pohľade je koncept zložený z množiny objektov, ktoré sú inštanciami konceptov domény a množiny atribútov, ktoré koncept opisujú. Pri intencionálnom pohľade je koncept zložený z množiny atribútov, ktoré ho opisujú. Vzájomným vzťahom medzi extencionálnym a intencionálnym pohľadom sa zaoberá formálna analýza konceptov (angl. Formal Concept Analysis). Keďže sa v poslednom čase čoraz častejšie objavujú extencionálne ontológie, vzniká potreba porovnávať inštancie konceptov takýchto ontológií.

V prípade, že používateľovi prezentujeme dva koncepty a on explicitne ohodnotí napr. záujem o ne, porovnaním a ďalším skúmaním spoločných a odlišných atribútov týchto konceptov môžeme získať zaujímavé informácie o používateľovi. Predpokladajme, že používateľ ohodnotí záujem o dva koncepty odlišnými

hodnotami (jeden pozitívne, druhý negatívne). Ak pri výpočte podobnosti zistíme, že sa koncepty významne odlišujú len v jednom atribúte, môžeme usúdiť, že tento atribút ovplyvnil hodnotenie používateľa. Podobne môžeme usudzovať o konceptoch, ktoré majú rovnaké ohodnotenie. V tomto prípade sú pre používateľa zrejme dôležité atribúty s rovnakou hodnotou.

V príspevku sa venujeme niektorým aspektom určenia podobnosti konceptov s cieľom zohľadniť jedinečnosť používateľa pri výpočte podobnosti.

2 Podobnosť inštancií konceptov

Z matematického hľadiska podobnosť dvoch objektov vyjadrujeme číslom z intervalu $\langle 0,1 \rangle$, pričom pri úplne odlišných objektoch sa podobnosť rovná nule a pri identických objektoch sa rovná 1. Charakteristikami podobnosti sú reflexia (objekt je identický sám so sebou) a symetria (ak sa podobá A na B , tak sa podobá aj B na A).

Vo všeobecnosti, ontologické koncepty sú zložené z atribútov, ktoré môžu byť dátové alebo objektové. Zároveň môžu byť niektoré atribúty násobné. Porovnanie hodnôt každého z atribútov prispieva k celkovej podobnosti inštancií, ktorá je daná súčtom týchto čiastkových podobností.

2.1 Zapojenie modelu používateľa

Jednoduchým spočítaním čiastkových podobností (alebo použitím inej agregáčnej funkcie) dostaneme pre dva koncepty vždy rovnaký výsledok za každých okolností (samozrejme pri použití rovnakých stratégií porovnávania). To nie je postačujúce v personalizovanom prostredí, kde máme k dispozícii model používateľa zložený z charakteristík, ktoré vyjadrujú jeho preferencie.

Riešením tohto problému je zavedenie princípu váh do výpočtu podobnosti. To, čo sa používateľovi páči (dané charakteristikami v modeli používateľa), by malo viac zavážiť. V princípe každá čiastková podobnosť má váhu rovnú 1, čo sa nijako neprejavuje vo výpočte. Preto čiastkovú podobnosť atribútu konceptu zdvojnásobíme, ak v modeli používateľa existuje zodpovedajúca charakteristika a jej hodnota sa rovná hodnote atribútu konceptu. V prípade, že sa v modeli používateľa nachádza zodpovedajúca charakteristika, ale jej hodnota nie je rovná hodnote atribútu konceptu, váha môže nadobúdať hodnotu z intervalu $(1,2)$.

Zavedením váh do určenia podobnosti chceme dosiahnuť zvýšenie vplyvu modelu používateľa vo výpočte pre konkrétne atribúty konceptu. Výška maximálnej váhy môže byť odlišná v závislosti od aplikačnej domény a množstva atribútov konceptu. Overenie, či navrhnutá dvojnásobná váha je dostatočná na ovplyvnenie celkovej podobnosti, treba vykonať experimentálne.

2.2 Pridanie váh prechádzaním inštalácie konceptu do hĺbky

Ukázali sme, ako charakteristiky modelu používateľa umožňujú zvýšiť podobnosť prostredníctvom váh. Podobne môžeme použiť váhy na zníženie podobnosti.

Ako sme spomenuli už skôr, koncept môže mať dátové a objektové vlastnosti, pričom na objektové vlastnosti sa môžu opakovane napájať ďalšie vlastnosti (aj objektové aj dátové), čím inštancia konceptu narastá do hĺbky. Ak by sme zachovali pôvodný princíp bez váh (váha rovná 1), prípadná nezhoda na najnižšej úrovni by ovplyvnila podobnosť rovnako ako nezhoda na najvyššej úrovni, čo nezodpovedá chápaniu podobnosti. Preto čiastkovú podobnosť vynásobíme váhou menšou ako 1, pričom pri postupnom prechádzaní na nižšiu úroveň sa bude váha znižovať, čím bude menej vplývať na celkovú podobnosť.

2.3 Podobnosť alebo príbuznosť?

Výslednú podobnosť sme navrhli ako súčet čiastkových podobností vydelenú počtom uskutočnených porovnaní (resp. počtom porovnávaných atribútov). Pri porovnávaní inštancií ontologických konceptov často nastáva situácia, keď porovnávané inštancie majú nielen odlišné hodnoty atribútov, prípadne ich počet, ale aj samotné atribúty sú odlišné. Ak má atribút viacnásobný výskyt v oboch inštanciách, ale s rozdielnym počtom, situáciu riešime tak, že vytvoríme dvojicu atribútov (podľa definovanej miery podobnosti) a tie porovnáme. Zvyšné atribúty vyhodnotíme, akoby nemali v druhej inštancii atribút na porovnanie.

V tomto prípade a v prípadoch, keď sa atribút v jednej z inštancií nenačádza, usudzujeme, že sú maximálne odlišné, a teda ich výsledná čiastočná podobnosť je nulová. Podľa [5] sémantická podobnosť je špeciálnym prípadom sémantickej príbuznosti (angl. relatedness). Teda v tomto špeciálnom prípade, keď inštancie nemajú rovnaké atribúty, by sme mali hovoriť už o sémantickej príbuznosti.

3 Podobnosť a charakteristiky používateľa

Z explicitného ohodnotenia, napr. záujmu používateľa o koncepty (pri používaní systému poskytujúceho informácie), môžeme usudzovať o charakteristikách používateľa. Ak koncept obsahuje atribút, ktorý je pre používateľa dôležitý, pravdepodobne ovplyvní jeho hodnotenie viac smerom k pozitívnym hodnotám a naopak, ak je atribút neprijateľný, smerom k negatívnym hodnotám.

Z tohto dôvodu zavádzame dve prahové hodnoty, ktoré rozdelia atribúty podľa vypočítanej podobnosti na tri množiny. Keďže nás zaujímajú najmä atribúty, ktoré výrazne vplývajú na celkovú podobnosť, prahové hodnoty nerozdelia interval podobnosti $\langle 0,1 \rangle$ na rovnaké intervaly. Pre zaradenie do množiny pozitívnych atribútov sme navrhli pri experimentoch podobnosť vyššiu ako 0,85 a pre zaradenie medzi negatívne atribúty podobnosť menšiu ako 0,15. Overenie týchto prahových hodnôt je predmetom ďalších experimentov.

Touto metódou predpripravené atribúty môžu iné nástroje transformovať na charakteristiky používateľa a použiť ich na naplnenie alebo v prípade existujúcich charakteristík na aktualizáciu modelu používateľa. Na základe dostatočne naplneného modelu a v danom čase aktuálneho modelu používateľa môže prebiehať prispôsobovanie prezentácie informácií a/alebo navigácie.

4 Záver

V príspevku navrhujeme prístupy k porovnávaniu inštancií ontologických konceptov s cieľom zohľadniť vo výpočte používateľovu jedinečnosť. Pre potreby personalizácie hľadáme pri porovnávaní jednotlivých atribútov dôvody, ktoré mohli spôsobiť explicitne ohodnotenie, napr. záujem používateľa o koncepty a vytvárame množiny pozitívnych a negatívnych atribútov. Pri takto získaných atribútoch môžeme usudzovať o charakteristikách používateľa a následne ich použiť na naplnenie modelu používateľa charakteristikami a tiež na aktualizáciu existujúcich charakteristík v modeli používateľa.

Takto predpripravené dáta môže využiť napríklad nástroj *Log Analyzer* [1] realizujúci automatické získavanie charakteristík používateľa sledovaním správania a analyzovaním záznamov o vykonaných akciách. Opísané prístupy plánujeme overiť na inštanciách ontológie pracovných ponúk zúženej na oblasť informatiky (vytvorená v rámci projektu NÁZOU [3]), na inštanciách ontológie publikácií (projekt MAPEKUS [2]) a odpovedajúcich modeloch používateľa.

Tento príspevok vznikol za podpory Štátneho programu výskumu a vývoja "Budovanie informačnej spoločnosti" č. úlohy 1025/04 a Vedeckej grantovej agentúry VEGA v rámci grantovej úlohy č. VG1/3102/06.

Referencie

1. Andrejko, A., Barla, M., Bieliková, M., Tvarožek, M.: User characteristics acquisition from logs with semantics. In ISIM '07 Information Systems and Formal Models: 10th International Conference on Information System Implementation and Modeling. (2007) 103–110
2. Bieliková, M., Návrat, P.: Modelovanie a získavanie, spracovanie a využívanie znalostí o konaní používateľa v hyperpriestore Internetu. In Proceedings of Znalosti 2007, Ostrava, Česká republika, Február (2007) 368–371
3. Bieliková, M., Návrat, P., Vojtáš, P., Hluchý, L., Bartoš, P.: Softvérové nástroje pre získavanie, organizovanie a prezentáciu pracovných ponúk na webe. In Proceedings Of Datakon 2006, Brno, Česká republika (2006) 1–20
4. Brusilovsky, P., Tasso, C.: Preface to Special Issue on User Modeling for Web Information Retrieval. *User Modeling and User-Adapted Interaction* 14 (2-3), (2004) 147–157
5. Budanitsky, A., Hirst, G.: Semantic distance in WordNet: An experimental application-oriented evaluation of five measures. In Workshop on WordNet and other lexical resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburg, June (2001)
6. Formica, A.: Ontology-based concept similarity in Formal Concept Analysis. *Information Sciences* (18), (2006) 2624–2641
7. Svátek, V.: Ontologie a WWW. In Datakon 2002. Brno, Česká republika (2002) 1–35