Tudásalapú ajánlórendszer adatszegény környezetben

Oravecz Csaba, Sárközy Csongor, Mittelholcz Iván

MTA Nyelvtudományi Intézet
e-mail:{oravecz.csaba,sarkozy.csongor,mittelholcz.ivan}@nytud.mta.hu

Kivonat Az ajánlórendszerek általában a felhasználói tranzakciókból és a termékekről rendelkezésre álló adatokból kinyert információkra támaszkodnak. Adatszegény környezetben azonban más információforrások felhasználására van szükség. A tanulmány olyan megoldás prototípusát mutatja be, ahol a felhasználó tevékenységét jellemző szöveges adatok automatikus feldolgozása és egy részletes ontológiában tárolt tudásbázis felhasználása segítségével válik lehetővé a releváns termékek (szolgáltatások) kiválasztása.[1]

Kulcsszavak: ajánlórendszer, tudásbázis, ontológia

1. Bevezetés

Az online felhasználók számára az igényeiknek megfelelő termékek és szolgáltatások keresése, azonosítása és beszerzése a kérdéses termék, szolgáltatás komplexitásának függvényében komoly kihívást jelentő feladat lehet, melyben a felhasználó megfelelő támogatása kritikus fontosságú. A legegyszerűbb keresőalkalmazások általában azzal a feltételezéssel működnek, hogy a felhasználó pontosan tisztában van az általa keresett termék releváns paramétereivel és kimerítően ismeri az elérhető termékek halmazát is. Ez azonban ritkán vagy így, ezért az adott döntési folyamatban helye van annak az (automatikus) asszisztenciának, amely képes a felhasználót hatékonyan segíteni.

Azokat az internetes alkalmazásokat, melyek a felhasználók érdeklődésére számot tartó termékek és szolgáltatások felderítésében és kiválasztásában nyújtanak automatikus segítséget, ajánlórendszereknek (*recommender systems*) [6] nevezzük. Esetünkben olyan ökoinnovációs intézkedések, szolgáltatások személyre szabott kiajánlását végzi egy automatikus rendszer[2], melyek alkalmazásával a felhasználók (vállalkozások) jelentős megtakarítást érhetnek el. Az alkalmazás felépítésében és funkciójában sok hasonlóságot mutat az ajánlórendszerek klasszikus típusaival, ezen belül is a tudás- és megszorításalapú rendszerekkel [1,7], de az általunk fejlesztett rendszer működési tartományának és környezetének egyúttal számos olyan paramétere van, melyek egyedileg kidolgozott megoldásokat követelnek meg, túlmutatva a klasszikus ajánlórendszerekben felhasznált módszerek és algoritmusok szolgai alkalmazásán.

[1] A kutatást a KMR_12-1-2012-0036 számú, *Piacorientált kutatás-fejlesztési tevékenység támogatása a közép-magyarországi régióban* pályázat támogatta.

[2] A továbbiakban ECOINNO rendszerként hivatkozott alkalmazás.

2. Kihívások az ECOINNO rendszerben

Alapvetően három szempont szerint érdemes megvizsgálni azokat a tulajdonságokat, melyek lényeges eltéréseket jelentenek a megszokott ajánlási paradigmához képest.

- Feladat és kontextus: Mind a kínálati, mind a keresleti oldalon speciális paramétereket kell figyelembe venni. A termékek igen komplex szolgáltatások, melyek releváns tulajdonságainak meghatározása és reprezentációja nem triviális, ennek a feladatnak automatikus, gépi módszerekkel történő megoldása nagy kihívást jelentő probléma. Másrészt a kínálati halmaz számossága nem akkora, hogy a humán szakértői beavatkozást eleve ki kellene zárni, vagyis a terméktulajdonságok (automatikusan segített) manuális annotációja reális alternatíva. Ez a megközelítés a későbbi esetleges kiterjesztés során is fenntartható, hiszen a kérdéses szolgáltatások várható jövőbeli bővülése messze nem olyan ütemű, ami a kézi feldolgozást lehetetlenné tenné.

  A keresleti oldalon megjelenő felhasználók többsége először kerül kapcsolatba a rendszerrel, illetve a rendszer által ajánlott terméktípussal. Egyrészt tehát gyakorlatilag minimális mértékben tudja explicit módon megfogalmazni a valós igényeit, másrészt kezdetben semmilyen információ nem áll rendelkezésre arra vonatkozóan, hogy korábban milyen hasonló termékeket vett igénybe. Ezen túl, a rendszer használata során sem várható olyan mennyiségű ilyen jellegű adat felhalmozódása adott felhasználóval kapcsolatban, melyre a további ajánlatokat alapozni lehetne, így ez a fajta információ csak minimális mértékben vehető figyelembe.

  Fontos szempont, hogy a rendszer alkalmazási területe jól meghatározott, zártnak tekinthető, ezért az ezzel kapcsolatos háttértudás egyértelmű meghatározása és formális rögzítése természetesen kínálkozó lehetőség.
- Információforrás: Mindkét oldalon igen változatosak a nyers, közvetlenül elérhető információ formai és tartalmi jellegzetességei.
  - Termékoldalról: A szolgáltatások strukturálatlan vagy félig strukturált[3] leírásai, melyek a legritkább esetben készültek azzal a céllal, hogy automatikus számítógépes módszerekkel feldolgozhatók, értelmezhetők legyenek.
  - Felhasználói oldalról: Minden olyan információforrás, mely a felhasználó tevékenységére, környezetére vonatkozóan tartalmaz adatot, és a rendszer számára a felhasználó minimális közreműködésével hozzáférhető, releváns lehet (weboldal URL, prospektusok, ismertetők, beszámolók, jelentések stb.). Ez a fajta információ alapvetően strukturálatlan szöveges

[3] Strukturált információ a rendszer szempontjából olyan formátumú adat, mely explicit, géppel értelmezhető formális reprezentációba közvetlenül, nyelvi, logikai feldolgozás nélkül leképezhető.

formában jelenik meg[4], és független a felhasználó és az ajánlórendszer közötti interakciótól, nem abból származik.
– Kívánt eredmény: A felhasználói igényeknek megfelelő szolgáltatások rangsorolt listája jelenik meg a rendszer kimeneteként[5].

3. A megvalósítás általános elvei

A rendszer működése során alapvetően egy információ-visszakeresési (*information retrieval (IR)*) problémát old meg, ahol a klasszikus alkotóelemek, mint a *dokumentumgyűjtemény* és a *keresési kifejezés* speciális formában jelennek meg. Ezért a megfelelő módosításokkal a IR-paradigma, illetve meghatározott sztenderd algoritmusok alkalmazása nyitva áll a feladat megoldása során. Ha kínálati oldalon elérhető termékek (mint „dokumentumok") és a keresleti oldalon megjelenő felhasználók igényei (mint „keresőkifejezés") alkalmas módon meghatározhatók és reprezentálhatók, a feladat ezen reprezentációk közötti hasonlóság kiszámítására redukálódik (mint klasszikus IR-probléma).

Két feladatot kell tehát megoldani. Egyrészt a felhasználói profil és a terméktulajdonságok olyan reprezentációját meghatározni, mely lehetővé teszi ezen reprezentációk között egy hasonlóságmérték definiálását és kiszámítását, másrészt a rendelkezésre álló információforrások adatait (és esetleg további információforrásokat) felhasználva mindkét oldalon előállítani ezeket a reprezentációkat. Nyilvánvaló, hogy az előzőekben említett zárt és korlátos domén lehetőséget ad arra, hogy a releváns háttérismeret, fogalmi rendszer és összefüggések egy formális explicit leírásban (ontológiában) megadhatóak legyenek [2,8], ennek a tudásbázisnak a felhasználása kritikus fontosságú.

4. Reprezentáció

A *kínálati oldalon* a reprezentációk előállítása a szolgáltatások deskriptív jellemzését tartalmazó szöveges leírások számítógépes nyelvészeti elemzését, felcímkézését, majd ezen címkézés manuálisan segített, a tudásbázis (ontológia) által definiált fogalmi térbe történő leképezését foglalja magában. A *keresleti oldalon* ugyanez történik azzal a különbséggel, hogy a folyamat teljesen automatikus, forrásként a felhasználóról elérhető minden lehetséges deskriptív adatot felhasznál, illetve támaszkodhat a felhasználótól irányított formában bekért további adatokra (preferenciákra).

A rendszer az alkalmazási doménre vonatkozó háttértudást, releváns objektumokat, kategóriákat, fogalmakat és relációkat egy explicit formális tudásbázisban (ontológiában) tárolja. Az ontológia az OWL Web Ontology Language

[4] Természetesen a szöveg főbb alkotóelemeire (cím, fejléc, bekezdés stb) tagolt és ebben az értelemben strukturált lehet, de ez a rendszer számára csak segédinformáció, mely további feldolgozásra vár annak érdekében, hogy a kivont információ a megelőző lábjegyzetben leírt értelemben is strukturált legyen.

[5] Ebből a szempontból az ECOINNO rendszer nem különbözik a sztenderd ajánlórendszerektől.

formalizmusban implementált, építésének munkakörnyezete a Protégé ontológiaszerkesztőre és a hozzá kapcsolódó Java API-ra épül[6].

Tegyük fel, hogy a rendszer alkalmazási tartományában lehetséges a releváns háttértudást „kimerítően" leírni.[7] Ebben az esetben mind a kínálati oldal intézkedései, mind a keresleti oldal felhasználói profiljai megfogalmazhatók, leírhatók a tudásbázis segítségével, mint olyan fogalomhalmazok, melynek elemei az ontológia adott csomópontjaihoz tartozó fogalmak. Alapvetően tehát mindkét oldalon a kérdéses reprezentáció egy többdimenziós fogalomvektor, ahol a vektor koordinátái az ontológia által specifikált fogalmak, értékei pedig kétfélék lehetnek: bináris vektor esetén 0 vagy 1, valós vektor esetén pedig az adott fogalom relevanciáját reprezentáló súlyérték. Bináris vektorok előállítása rendkívül egyszerű, amennyiben az adott fogalom hozzárendelhető az intézkedés, illetve felhasználói profilhoz, az érték 1, egyébként 0. Valós vektor esetén meg kell határozni azt a módszert, melynek segítségével az értékek kiszámíthatók.

4.1. Az intézkedésprofil előállítása

Mivel a kiajánlható szolgáltatások halmazának számossága nem kirívóan nagy, elvileg a teljesen manuális annotáció sem kivitelezhetetlen. Célszerű azonban az annotációs folyamatot automatikus módszerekkel segíteni. Ekkor a szolgáltatásokról rendelkezésre álló szöveges információt egy nyelvi elemzőlánc dolgozza fel és annotálja egy előre definiált címkehalmazból (amely az ontológia alacsony szintű specifikus fogalmainak feleltethető meg) választott címkékkel. A humán annotátor ezek után ellenőrzi és javítja a hozzárendelést, illetve a kezdeti változatban súlyokat rendel a megfelelő címkékhez. Az így kialakított reprezentáció a 4.3. részben ismertetett módon kerül további feldolgozásra.

4.2. A felhasználói profil előállítása

A feladat a felhasználói oldalon elérhető jórészt strukturálatlan adatokból az ajánlórendszer számára releváns információ kinyerése és leképezése a tudásbázis által specifikált vektortérbe. Ez több lépésben történik.

– Forrás-előfeldolgozás. Első lépés a szöveges adatok típusától függően (pl. HTML-dokumentum, PDF-ismertető, Word-dokumentum stb.) a dokumentumstruktúra elemeinek azonosítása (keletkezési idő, cím, fejléc, bekezdés stb.), mely lehetővé teszi az információ pontos lokalizálását (ezáltal pl. súlyozását) az adott forráson belül.
– Információazonosítás. Ebben a lépésben a szöveges adatok nyelvi, szemantikai elemzésére kerül sor, ahol részletes annotációt kapnak a tartalmas nyelvi elemek és szerkezetek, megtörténik a kulcskifejezések és a köztük lévő viszonyok meghatározása.

[6] `http://protege.stanford.edu/overview/protege-owl.html`

[7] Nyilván ezt közvetlenül mérni nem lehetséges, a tudásbázis minőségére gyakorlati szempontból a rendszer működési hatékonyságából, pontosságából lehet következtetni.

– Leképezés. Jelen specifikáció szerint a doménontológia alsó szintű fogalmaihoz vannak hozzárendelve azok a nyelvi elemek és relációk, melyek a kérdéses fogalmakat a szöveges adatokban instanciálják. Alapvetően tehát a tudásbázis definiálja a felhasználói adatok annotációjából az ontológiai fogalmakba történő leképezést. Mind a fogalmi csomópontok, mind a releváns nyelvi elemek azonosításában nagy szerepet játszanak azok a lexikális erőforrások, melyeket a rendelkezésre álló szöveges adatokból sztenderd statisztikai eljárások segítségével készültek (lásd 1. és 2. ábra).

```
1993.55969884557        1356    hulladék
1553.50432823931        1087    környezeti
1529.70377199569        1104    megtakarítás
1385.73629885459        1115    intézkedés
1068.88348017766        786     tonna
1065.25227892306        824     környezetvédelmi
895.661275426298        730     kft
709.212863548727        639     beruházás
675.285411574036        608     termék
...
```

1. ábra. Felhasználói dokumentumokból előállított kulcsszólista.

```
nyers<>fűrészpor<>10 12.2717 3.1616 10 14 34
kompakt<>fénycső<>15 11.7500 3.4631 12 20 41
szabadlevegős<>hűtés<>16 11.7395 3.4631 12 14 59
szerves<>oldószer<>32 10.7642 3.8708 15 29 70
mart<>aszfalt<>15 12.7571 3.1618 10 10 34
hulladékhő<>hasznosítás<>44 9.9866 3.1592 10 16 145
épület<>fűtés<>73 8.5454 3.4548 12 108 70
maradékanyag<>mennyiség<>74 8.5424 3.3077 11 16 434
...
```

2. ábra. Felhasználói dokumentumokból előállított kollokációs lista.

A felhasználói profilvektorhoz súlyokat hozzárendelni legegyszerűbben instanciagyakoriság alapján lehet:

$$w_{i,j} = \frac{freq_{i,j}}{\max_k freq_{k,j}} \tag{1}$$

ahol $w_{i,j}$ a j felhasználóprofilban instanciálódott i fogalomhoz tartozó súlyérték, $freq_{i,j}$ i előfordulási gyakorisága j-ben, $\max_k freq_{k,j}$ pedig a leggyakoribb fogalomhoz tartozó gyakorisági érték. A reprezentációkat az 1. táblázat illusztrálja.

1. táblázat. Profilreprezentációk.

	...	C_n(szennyvíz)	C_{n+1}(talaj)	C_{n+2}(zaj)	...
intézkedés$_i$	...	0.023	0.001	0	...
intézkedés$_j$	...	0	0.001	0.145	...
intézkedés$_k$	...	0	0.326	0.002	...
			...		
vállalkozás$_i$	...	0.001	0.001	0	...
vállalkozás$_j$	...	0.423	0.003	0	...
vállalkozás$_k$	...	0.003	0.377	0.005	...

4.3. A fogalmi vektorok kiterjesztése

A fogalmak közötti viszonyokat explicit módon specifikáló doménontológia lehetőséget ad arra, hogy a közvetlenül instanciált fogalmak mellett a velük meghatározott módon kapcsolatban álló további csomópontok (fogalmak, fogalmi osztályok) is hozzáadódjanak a reprezentációs vektorhoz. Ez a kiterjesztés például az ún. megszorított terjedésaktiváció (*constrained spreading activation*) alkalmazásával valósítható meg [3,4], melynek során különböző megszorítások által korlátozott módon az egyes csomópontokhoz további kapcsolódó csomópontok rendelhetők hozzá, ily módon az eredeti fogalomvektor kibővül a kapcsolódó fogalmakkal.[8]

A profilok választott vektortér alapú reprezentációja lehetővé teszi, hogy sztenderd hasonlósági mértékek segítségével természetes módon rangsorolhatók legyenek a felhasználókra szabott ajánlatok. A jelenleg alkalmazott mérték a koszinusz hasonlóság.

5. Javító stratégiák

A rendszer kézenfekvő kiterjesztését adják olyan szabályalapú megszorítások, melyek mind a felhasználói profil nyelvi szemantikai elemzéséből származó annotáció, mind az intézkedések tulajdonságai, mind a doménontológia szintjén megfogalmazhatók, és bizonyos kapcsolatokat, következményeket egyértelműen definiálnak[9]. További hasonló megszorítások származtathatók az audit kérdőívekre adott felhasználói válaszokból. Ezek egyértelmű és kiterjedt specifikálása a rendszer szempontjából kritikus fontosságú, mivel nagy mértékben leszűkíthetik az illesztési probléma keresési terét, javítva az ajánlati válaszlistát.

Nincs olyan mesterséges intelligenciára támaszkodó alkalmazás, amely kimerítően képes lenne kezelni egy adott tárgykört, tartományt. Előfordulhat, hogy

[8] A kapcsolt fogalmakhoz rendelt súly számítható pl. az eredeti súlyból lépésenként konstans érték levonásával (*decay*).

[9] Pl. adott tulajdonsággal rendelkező, vagyis adott dimenzióban nem 0 értékű vektorral jellemzett intézkedés nem járhat együtt egy másik meghatározott módon specifikált intézkedéssel.

a rendszer tudásbázisa hiányos, és nem lehet megbízható profilt, reprezentációt előállítani a felhasználóról a rendelkezésre álló adatok alapján. Ilyenkor lehetőség van arra, hogy a rendszer alacsonyabb szintű, kulcsszóalapú illesztést végezzen, illetve a tudásalapú és a kulcsszóalapú illesztést kombinálja. Ennek a megoldásnak a pontos paraméterei a részletes tesztelés során határozhatók meg.

6. Kiértékelési módszerek

Az ajánlórendszerek kiértékelésére nincs egységes, minden feladatban megbízhatóan alkalmazható módszertan [5]. Mind a tesztadatok kiválasztására, mind a felhasználói visszajelzésekből származó információ felhasználására számos megoldás lehetséges, ahol a konkrét alkalmazás teljesítményét legpontosabban mérő eljárás kidolgozása nem triviális. A projekt jelenlegi szakaszában egy mintegy 300 intézkedést tartalmazó tesztadatbázisra és 10-15 felhasználó bináris szelekciót tartalmazó válaszaira támaszkodik a kiértékelés[10]. Ebben a kontextusban a sztenderd fedés, pontosság értékek értelmezhetők, a tesztelési folyamat keresztvalidációval elvégezhető. Ez a megközelítés azonban meglehetősen durva modelljét adja a felhasználói elégedettségnek, ezért nem tekinthető végleges megoldásnak.

7. Összefoglalás és további feladatok

A tanulmányban bemutattuk egy olyan ajánlórendszer prototípusát, mely alkalmazási tartományának jellegéből nem rendelkezik azzal a jelentős méretű adathalmazzal, melyre a klasszikus rendszerek általában támaszkodnak, így a működéshez szükséges információt, tudást a sztenderd módszerektől eltérő úton kell megszerezni, illetve előállítani. Mint nagyon sok valós környezetben használt nyelvtechnológiai alkalmazás, az ECOINNO rendszer is hibrid megoldásokat alkalmaz, illeszkedve a feladat peremfeltételeihez. A rendszer alapvetően a tudásalapú megközelítéshez áll közel, de nem zárja ki más megközelítések kedvező tulajdonságainak kihasználását (pl. a felhasználói visszacsatolások figyelembevétele, elegendő felhalmozott adat esetén a felhasználóprofilok hasonlóságának monitorozása stb.).

Hivatkozások

1. Burke, R. Knowledge-based recommender systems. In: Kent, A. szerk.: Encyclopedia of Library and Information Systems, 69. kötet. New York, Marcel Dekker (2000)
2. Castells, P., Fernández, M., Vallet, D. An Adaptation of the Vector-Space Model for Ontology-based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering, Special Issue on "Knowledge and Data Engineering in the Semantic Web Era", **19**(2) (2007) 261–272

[10] A tudásbázis folyamatos fejlesztése miatt publikus adatok még nem állnak rendelkezésre.

3. Crestani, F., Lee, P. L. Searching the web by constrained spreading activation. Information Processing & Management, **36**(4) (2000) 585–605
4. Griffith, J., O'Riordan, C., Sorensen, H. A Constrained Spreading Activation Approach to Collaborative Filtering. In: Gabrys, B., Howlett, R.J., Jain, L.C. szerk. Knowledge-Based Intelligent Information and Engineering Systems. Lecture Notes in Computer Science, 4253. kötet. Berlin, Heidelberg, Springer (2006) 766–773
5. Gunawardana, A., Shani, G. A survey of accuracy evaluation metrics of recommendation tasks. The Journal of Machine Learning Research, **10** (2009) 2935–2962
6. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. szerk. Recommender Systems Handbook. Springer (2011)
7. Thompson, M., Göker, C., Langley, P. A Personalized System for Conversational Recommendations. Journal of Artificial Intelligence Research, **21** (2004) 393–428
8. Vallet, D., Fernández, M., Castells, P. An Ontology-based Information Retrieval Model. In: Proceedings of the 2nd European Semantic Web Conference (ESWC 2005), Heraklion, Greece (2005) 455–470