Kategorizácia príspevkov na blog-portáli za pomoci ontológie

Juraj Frank a **Martin Homola**

Univerzita Komenského, Fakulta Matematiky, Fyziky a Informatiky
Mlynská dolina, Bratislava.
juraj.frank@gmail.com, homola@fmph.uniba.sk

Abstrakt

Na portáli blog.matfyz.sk používame tagovanie ako základnú metódu kategorizácie blogových príspevkov. Tagovanie znamená, že autori priradia každému príspevku jedno alebo viac kľúčových slov, ktoré ho charakterizujú. S tagovaním sa spája niekoľko problémov. Rôzni autori použijú rôzne tagy pre podobné témy, a pod. Navrhujeme použiť pri kategorizácii blogových príspevkov ontológie a sémantické technológie. Ontologické dáta získame metódou dolovania ontológií z textu, aplikovanou na text príspevkov ako aj na autorom priradené tagy. Veríme, že takýto systém umožní automatickú sugesciu a korekciu v procese tagovania príspevku autormi, tematické klastrovanie príspevkov, a zjednoduší pre používateľov proces vyhľadávania článkov podľa ich konkrétnych preferencií.

1 Úvod

Socio-ekonomický prínos blogovania zaznamenáva v súčasnosti rapídny nárast. Komerčne zamerané blogy dokážu predávať produkty a otvárajú nové dvere vo sfére vzťahov so zákazníkom. Internetové média prevádzkujú blog-portály, prostredníctvom ktorých bežní používatelia internetu vytvárajú obsah bez nároku na odmenu. Tento obsah sa následne ponúka ostatným používateľom a majiteľ portálu zarobí na online reklame alebo pritiahnutím pozornosti na iné, komerčnejšie zamerané portály.

Milióny blogerov vyprodukujú denne kvantá príspevkov, pričom je zrejmé, že jeden konkrétny príspevok je zaujímavý len pre zlomok z nich. V sfére záujmu konkrétneho konzumenta je iba malý fragment z obrovského množstva informácií, ktoré každý deň pribudnú na internete. Vyhľadávanie zaujímavých článkov v tejto obrovskej množine je zaujímavým a netriviálnym výskumným problémom. Správna selekcia je totiž výrazne podmienená konkrétnymi osobnými záujmami a môže byť úplne rozličná v prípade dvoch rôznych užívateľov.

Základným a dnes už de facto štandardným prístupom ku kategorizácii obsahu na internete je tagovanie. Pri tagovaní je každej jednotkovej časti obsahu, či už ide o článok, obrázok, či napr. video, priradená množina kľúčových slov, ktoré nazývame tagy. Tagovanie používame aj na našom portáli blog.matfyz.sk. Tento prístup umožňuje používateľom portálu pristupovať k obsahu podľa témy. Keďže tagy sú priradené autormi textu, či inými používateľmi portálu, obsahujú v sebe „ľudskú perspektívu“, a preto vo všeobecnosti aj pre ostatných používateľov predstavujú intuitívnu pomôcku v prístupe k informáciám.

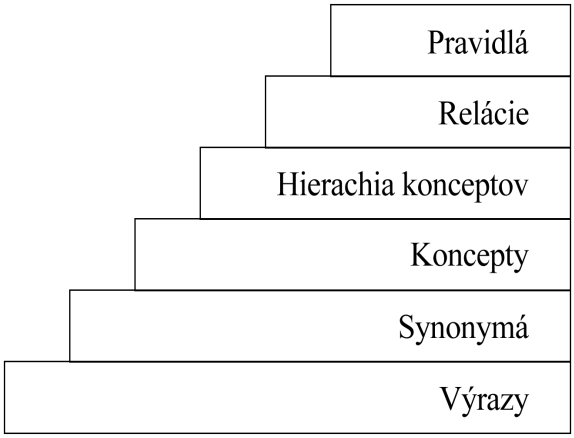
Motivácia. S tagovaním sa spája niekoľko problémov. Rôzni autori môžu napríklad použiť rôzne slová pre pomenovanie tých istých tém. Tiež sa stáva, že autor pri tagovaní minie niekoľko tém, ktoré tiež s článkom súvisia, a preto je pravdepodobné, že ich návštevníci portálu použijú pri vyhľadávaní obsahu. Sugescia tagov, založená na analýze textu príspevku by mohla napomôcť vyhnúť sa týmto problémom. Užitočnú podpornú sugesciu tagov môžeme tiež získať spomedzi tagov zo starších príspevkov, ktoré sú s novým príspevkom podobné. Klasickým prístupom k tomuto problému by bolo aplikovať kategorizáciu textov [10] a klastrovanie [9] založené na strojovom učení. V súčasnosti však aj sémantické technológie sú jedným z prístupov, ktorý bol aplikovaný v kontexte blogovania [8] a sociálnych sietí [7] a tento prístup sme sa rozhodli aplikovať aj my.

Náš prístup. Navrhujeme problém správy tagov na blog-portáli riešiť aplikáciou učenia sa ontológií z textu [1]. Kým klasické ontologické učenie pracuje s jediným vstupom, ktorým je text, my zapojíme ako vstup ešte aj tagy navrhnuté autorom. Nad textovým obsahom portálu budeme postupne budovať ontológiu, ktorú budeme rozširovať o nové koncepty vždy keď niektorý používateľ navrhne tagy pre novovytvorený príspevok. V našej ontológii sú teda koncepty úzko späté s tagmi. Zároveň použijeme metódy ontologického učenia na extrakciu ontologických relácií medzi takto získanými konceptami, tj. predovšetkým relácií subsumcie, ekvivalencie, partonómie, ale aj iných. Získanú ontológiu následne použijeme v procese sugescie tagov.

Očakávame nasledovné prínosy: zlepšenie tematického prístupu k článkom, zlepšenú kvalitu selekcie tagov autorom vďaka automatickej sugescii, objavovanie podobných článkov vďaka extrahovaným ontologickým reláciám medzi tagmi. Dúfame tiež v isté zlepšenie výkonu ontologického učenia, keďže naša aplikácia vo veľkej miere zapája do procesu ľudských používateľov portálu. Je nutné poznamenať, že náš prístup je v súčasnosti vo fáze návrhu a je potrebné ho preveriť implementáciou. Na tejto implementácii v súčasnosti pracujeme.

2 Učenie sa ontológie z textu

Pri učení sa ontológie z textu ide hlavne o extrahovanie konceptov a relácií z textu. Výskum v tejto oblasti je dôkladne opísaný v [1], v tejto sekcii poskytneme krátke zhrnutie. Úlohu ontologického učenia dosahujeme najčastejšie pomocou štatistických a lingvistických metód. Extrahovať potrebujeme však aj taxonomické a netaxonomické relácie medzi konceptami. Celková úloha je teda rozdelená do menších podúloh, ako napr. extrakcia výrazov a synoným, extrakcia konceptov, extrakcia taxonómie, relácií a pravidiel. Systém postupne zložitejších podúloh sa obvykle uvádza ako vrstvový koláč pre učenie sa ontológií [1] (viď Obrázok 1). O jednotlivých úlohách budeme hovoriť podrobnejšie v nasledovnom.



Obrázok 1: Vrstvový koláč pre učenie sa ontológií.

Extrakcia výrazov. Prvým krokom pri učení sa ontológie je extrakcia výrazov. Obvykle sa v texte označia slovné druhy a vzťahy susediacich a príbuzných slov. Označené podstatné mená a menné frázy sa následne použijú na konštrukciu ad-hoc lexikálno-syntaktických Hearst vzorov [6], ktoré odhaľujú vhodné výrazy. Takýmito vzormi môžu byť napríklad:

- <concept> such as <instance>,
- such <concept> as <instance>,
- <concept>, (especially|including) <instance>,
- <instance> (and|or) other <concept>.

Používame heuristiku *head matching* [11] (viď extrakcia relácií) alebo aj hlbšiu lingvistickú analýzu, založenú na technikách spracovania prirodzeného jazyka. Pre identifikáciu príbuzných výrazov do procesu vstupuje dodatočné štatistické spracovanie, ktoré porovnáva distribúciu výrazov v texte.

Extrakcia synoným. Pre získanie sémantických variácií výrazov musíme vykonať extrakciu synoným. Používajú sa buď už dostupné lexikálne zdroje (ako napr. WordNet[1]) alebo sa synonymá získavajú dynamicky pomocou klasteringu a podobných techník, postavených na distribučnej hypotéze, ktorá hovorí, že výrazy sú si významovo podobné do rovnakej miery, do akej zdieľajú syntaktické kontexty [5].

Extrakcia konceptov. Koncepty sú obvykle identifikované klastrami príbuzných výrazov. Extrakcia konceptov sa preto do veľkej miery prekrýva s extrakciou výrazov a synoným. Na extrakciu konceptov sa však možno pozerať z extenzionálneho alebo intenzionálneho uhla pohľadu. Extenzionálny prístup hovorí o definovaní konceptu pomocou objavovania jeho všetkých možných inštancií. To sa deje buď tak, že simultánne derivujeme koncept a jeho extenziu alebo napĺňame existujúce koncepty inštanciami. Naopak intenzionálny prístup sa zameriava na získavanie formálnych prípadne neformálnych definícií, pri neformálnych definíciách ide však skôr o výnimky.

Extrakcia hierarchie. Na extrakciu hierarchických vzťahov sa používa niekoľko metód, pričom niektoré z nich už boli spomenuté. Prvou z nich je použitie lexikálno-syntaktických Hearst vzorov [6] na objavenie asociatívnych výrazov. Takéto vzory sa však v texte vyskytujú veľmi zriedka. Hierarchickú reláciu možno indukovať pomocou metódy *head matching* [11], pri ktorej sa extrahuje relácia podtriedy medzi podstatným menom a tým istým podstatným menom s jeho modifikátormi (napr. „medzinárodná konferencia“ je podtriedou triedy „konferencia“). Iná technika využíva hierarchické klastrovacie algoritmy založené na Harrisovej distribučnej hypotéze. Medzi ne patrí formálna konceptová analýza, hierarchické klastrovanie, ale aj metóda k-priemerov [2].

[1] WordNet – lexikiálna databáza pre angličtinu: `http://wordnet.princeton.edu/`

Extrakcia relácií. Objavovanie štruktúry a závislostí pre extrakciu relácií sa pri väčšine prístupov deje pomocou štatistickej analýzy v kombinácii s lingvistickou analýzou rôznej zložitosti. Problémy riešené pri extrahovaní relácií sú navyše veľmi podobné problémom skúmaným pri spracovaní prirodzeného jazyka.

Extrakcia pravidiel. Niektoré fakty nemusia byť v ontológii uvedené explicitne, môžu sa však z danej ontológie dať odvodiť. Aby sme umožnili takéto odvodzovanie, mali by sme definovať, či dokonca objavovať odvodzovacie pravidlá. Toto je však pravdepodobne najmenej skúmaná oblasť učenia sa ontológií. Existujú však snahy zamerané na problémy lexikálneho usudzovania ako napr. [4], a preto je možné, že sa v blízkej budúcnosti objavia nové prístupy k učeniu sa ontologických pravidiel.

V oblasti učenia sa ontológií z textu sa doposiaľ urobilo veľa výskumu. Pre porovávanie rôznych prístupov je však dôležité dohodnúť sa na presnej úlohe tejto oblasti výskumu a vytvoriť náležité hodnotiace metódy. S takýmito univerzálnymi metódami hodnotenia bude možné porovnávať výkon rôznych prístupov na tej istej úlohe. Následne je možné očakávať zvýšený záujem a rast výskumu v tejto oblasti. Pre podrobnejšie informácie odkazujeme čitateľa na [1].

3 Budovanie ontológie z tagov s pomocou užívateľov

V tejto časti vysvetlíme, ako bude budovaná ontológia nad tagmi poskytnutými užívateľmi. Tento proces je opakujúci sa a zjemňovanie ontológie sa deje vždy, keď je publikovaný nový príspevok a prebehne fáza selekcie tagov. Tagy sú volené užívateľmi, ktorí zvažujú návrhy od systému. Posledné slovo voľby tagov majú vždy užívatelia, a teda ľudský pohľad a ľudská práca je v tomto procese zastúpená vo výraznej miere. Proces učenia sa ontológie popri priraďovaní tagov k blog-príspevkom teraz opíšeme podrobnejšie. Keďže budeme kombinovať niekoľko metód podobne, ako je to navrhnuté v [3], celý process sa bude odohrávat v niekoľkých fázach (viď Obrázok 2).

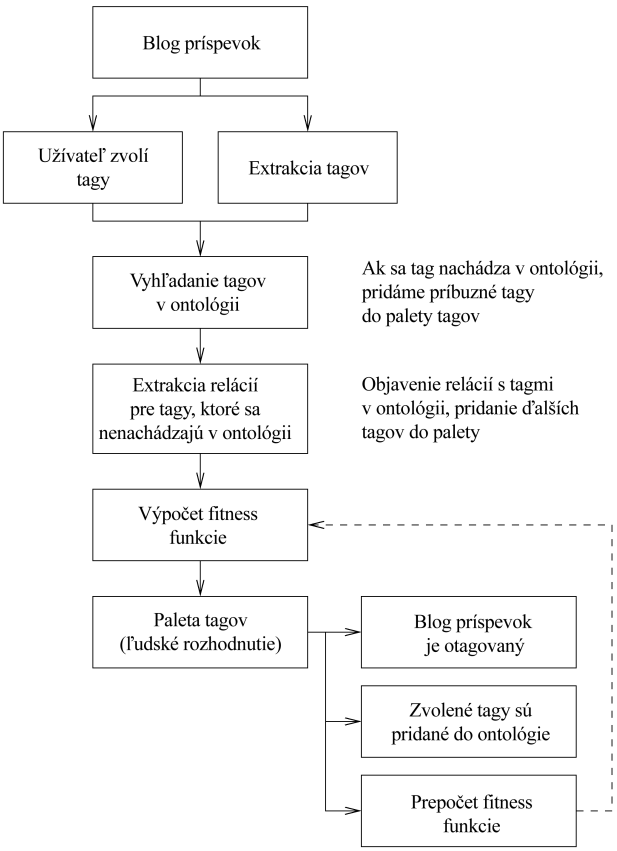
Extrakcia tagov. Autora nového príspevku do blogu požiadame po napísaní textu príspevku o priradenie tagov popisujúcich jeho obsah. Tieto tagy sú volené človekom, a teda mali by poskytovať dobrú základnú množinu tagov pre ďalšie spracovanie. V niektorých prípadoch sa však môže stať, že autor neposkytne žiadne tagy. Keďže v našom prístupe chápeme tagy ako koncepty, v takomto prípade tagy získavame pomocou algoritmov na extrakciu konceptov, ktoré sme spomenuli v druhej časti. Tieto algoritmy môžu byť využité aj na navrhutie vhodných tagov autorovi po zapísaní textu, či dokonca ich navrhovanie už počas samotného písania textu príspevku.

Vyhľadanie tagu v ontológii. V tomto kroku zisťujeme, či sa navrhovaný tag nachádza v našej portálovej ontológii alebo nie. Ak sa tag v ontológii nachádza, pridáme ho do skupiny tagov, ktorú budeme nazývať paleta tagov, túto neskôr ponúkneme autorovi (viď nižšie). Následne hľadáme v ontológii nadkoncepty, podkoncepty k danému konceptu, prípadne iné koncepty, ktoré sú s týmto tagom v nejakej netaxonomickej relácii. Ak takéto koncepty nájdeme, pridáme zodpovedajúce tagy do palety.

Extrakcia relácií. Ak sa niektorý z užívateľom navrhovaných tagov nenachádza v našej ontológii, chceme zistiť akými reláciami je spojený s konceptami, ktoré sa už v ontológii nachádzajú. Aby sme to dosiahli, navrhujeme tieto postupné kroky:

1. *Učenie sa relácií individuálov s pomocou Wikipedie.* Ak navrhovaný tag začína veľkým začiatočným písmenom, obvykle to značí, že môže ísť o inštanciu konceptu alebo individuál. Veľa z takýchto mien možno nájsť definovaných na Wikipedii[2]. Tento algoritmus vyhľadá definíciu, označí slovné druhy a odvodí všeobecný koncept, napr. vyhľadávanie mena Michael Jordan vráti definíciu „Michael Jordan is a basketball player“, z ktorej je následne odvodený koncept „basketball player“. Ak navrhovaný tag nezačína veľkým začiatočným písmenom alebo nenájdeme žiadnu definíciu, pokračujeme ďalším krokom.
2. *Head matching pre viacslovné tagy.* Ak má navrhovaný tag viac slov, spustíme algoritmus hľadania relácií založený na technike *head matching*. To umožní získať relácie ako napr. basketball player is-a player alebo dining room is-a room. Ak je navrhovaný tag iba jedno slovo, pokračujeme ďalším krokom.
3. *Učenie sa relácií s pomocou WordNet.* Vyhľadáme definíciu konceptu v sémantickom lexikóne WordNet a extrahujeme náležitý nadkoncept. Ak sa v lexikóne nenachádza definícia, pokračujeme ďalším krokom.

[2] Wikipédia – slobodná orvorená encyklopédia: http://en.wikipedia.org/ .



Obrázok 2: Schéma navrhovaného algoritmu.

4. *Učenie sa relácií všeobecných konceptov s pomocou Wikipedie.* Tento algoritmus sme už popísali v kroku č. 1, no namiesto inštancií a mien individuálov budeme vyhľadávať všeobecné koncepty, ktorých definíciu sme nenašli ani v lexikóne WordNet. Ak nenájdeme definíciu ani vo Wikipedii, pokračujeme ďalším krokom.
5. *Extrakcia relácií, založená na analýze textu príspevku.* Ak zlyhajú všetky predošlé kroky, na extrakciu relácií sú použité lexikálno-syntaktické vzory, prípadne klastrovacie algoritmy založené na Harrisovej distribučnej hypotéze. Vďaka tejto technike je možné objaviť nové koncepty, lokálne špecifické alebo komunitne špecifické koncepty.

Po vykonaní každého z týchto krokov zisťujeme, či sme sa naučili nejaké relácie medzi novým tagom a tagmi obsiahnutými v našej ontológii. Ak áno, nový tag pridáme do palety tagov spolu s tagmi, s ktorými je v relácii.

Výpočet fitness funkcie. Každému tagu nachádzajúcemu sa v palete tagov priradíme číselnú hodnotu pomocou fitness funkcie. Táto hodnota by mala vyjadrovať vhodnosť alebo schopnosť slova byť dobrým tagom. Bude počítaná z predchádzajúcich reakcií užívateľov portálu pri výbere tagov, pričom bude budovať na predpoklade, že určité slovo je vhodné alebo schopné byť dobrým tagom vtedy, ak ho užívatelia obvykle vyberú z palety tagov, naopak slová, ktoré obvykle nie sú vyberané z palety, budú zrejme nevhodné pre účely tagovania. Tagy, ktoré neboli použité nikdy predtým, budú mať priradenú nejakú základnú hodnotu, ktorá môže byť taktiež pozmenená podľa hodnôt fitness funkcie príbuzných tagov, objavených počas fázy učenia sa relácií.

Ľudské rozhodnutie. Tagy obsiahnuté v palete tagov ponúkneme autorovi cez užívateľské rozhranie takým spôsobom, ktorý berie do úvahy hodnotu fitness funkcie. To docielime buď upravením veľkosti písma pre každý tag, úpravou poradia tagov alebo aj nezobrazovaním tagov, ktoré nedosiahnu istú najnižšiu prahovú hodnotu fitness funkcie. Autor napokon spraví finálne rozhodnutie o priradení tagov k príspevku zvolením jedného alebo viacerých ponúknutých tagov. Ak sú niektoré z týchto tagov nové pre našu ontológiu, pridáme ich do nej. Výstupom tohto kroku bude:

Obrázok 2: Schéma navrhovaného algoritmu.

- blog príspevok, ktorý bude otagovaný aj ďalšími tagmi vďaka návrhom tagov, ktoré sú založené na ontologických vzťahoch,
- obohatenie ontológie tagmi navrhnutými užívateľom,
- prepočítanie fitness funkcie.

Navrhovaný algoritmus umožňuje pomerne priamočiare budovanie ontológie, pričom zároveň využíva už existujúce ontologické štruktúry. To sa deje vďaka využívaniu precíznych algoritmov ako napr. head matching či externých zdrojov ako napr. sémantických definícií nadkonceptov z lexikónu WordNet alebo článkov z Wikipedie. Filtrovanie všetkých možných tagov cez metriku fitness funkcie umožňuje, aby boli užívateľovi ponúknuté iba niektoré dôležité koncepty, čo uľahčuje celkový proces učenia sa ontológie s využitím tagov.

4 Očakávané výsledky

Hlavným cieľom našej práce je začať využívať ontológie a sémantickú technológiu na našom blog portáli. Aby sme poskytli čitateľom možnosť pristupovať k obsahu podľa témy, používame v súčasnosti tagovanie príspevkov tak, ako je to obvyklé aj pri iných komunitných portáloch. Tagovanie však so sebou prináša niektoré nevýhody ako napr.:

- Rôzni užívatelia používaju rôzne tagy na popis podobných tém. Toto je vskutku prirodzené, keďže aj prirodzený jazyk je často nejednoznačný.
- Chybná voľba tagov. Málokedy sa stáva, že užívatelia si vyberú tagy, ktoré nemajú žiadne spojenie s príspevkom. Omnoho častejšie sa stáva, že užívatelia volia tagy, ktoré majú s príspevkom len veľmi okrajový súvis.
- Neschopnosť poskytnúť tag pre významnú tému obsiahnutú v príspevku. Stáva sa to, keď sa užívatelia koncentrujú na určitú tému a neuvedomujú si, že výsledný príspevok zahŕňa aj iné témy.

Aj keď tieto problémy môžu byť riešené mnohými inými prístupmi, veríme, že usporiadanie tagov v ontológii, ktorú získavame učením sa z obsahu nášho portálu za pomoci kolaboratívneho editovania/spätnej väzby tak, ako sme navrhli v predchádzajúcej časti, nám značne pomôže tieto problémy zvládať. Predpokladáme hlavne tieto výhody a vylepšenia:

- *Zlepšenie prístupu užívateľov k obsahu* vďaka grupovaniu príbuzných tagov v prehliadači tagov. Zároveň plánujeme využiť subsumčnú hierarchiu ontológie tak, aby sme mohli utriediť tagy do prezerateľnej stromovo-hierarchickej štruktúry. Táto štruktúra bude prezerateľná aj podľa iných zaujímavých ontologických relácií ako napr. partOf, oppositeOpinionOf, a pod.
- *Pomoc používateľom voliť správne tagy* prostredníctvom návrhov vhodných tagov a overovaním zvolených tagov založenom na analýze textu s existujúcou ontológiou na pozadí.
- *Nájdenie príbuzných príspevkov* vďaka reláciám medzi tagmi zaznamenanými v ontológii. Príspevky s podobnou témou budú odporúčané čitateľom. Pre ilustráciu možných výhod zvažujme príspevok, ktorý má priradené tagy „concert“, „orchestra“ a „conductor“. Ak sú všetky tieto tagy príbuzné ontologickými reláciami k tagu „classical music“, môžme ponúknuť čitateľovi iné príbuzné príspevky označené len tagom „classical music“. Iný prípad by nastal, keby príspevok pokrýval nejakú kontroverznú tému. Dúfame, že vďaka našej ontológii budeme schopní objaviť a označiť opačné názory. Pre ilustráciu zvážme dva príspevky, pričom jeden má tag „liberal Opinion“ a druhý má tag „conservative Opinion“ a oba tieto tagy sú si v ontológii príbuzné cez rolu oppositeOpinionOf.
- *Vylepšenie učenia sa ontológie* pomocou kolaboratívneho editovania užívateľmi portálu. Náš portál má editorov, ktorí budú korigovať výsledky učiaceho algoritmu v prípade chyby a zároveň budú naďalej zjemňovať ontológiu, napr. keď učiaci algoritmus zabudne na niektoré relácie alebo jednoducho len textový korpus nebude dostatočne bohatý. Nesmieme zabudnúť, že náš systém využíva aj ľudskú inteligenciu, keďže zvažuje aj tagy navrhnuté priamo užívateľom.

Veríme, že s týmito výhodami môže náš prístup prispieť ako do oblasti komunitných portálov, tak aj do oblasti učenia sa ontológií. Keďže k tomuto problému existujú aj iné prístupy, bude zaujímave porovnávať naše výsledky s inými metódami. Zaujímať nás bude predovšetkým porovnanie s čisto štatistickými sub-symbolickými prístupmi, ako aj s čisto kolaboratívnymi prístupmi, ktoré sú postavené len na ľudskej práci.

5 Záver

Popísali sme návrh prístupu, ktorý plánujeme aplikovať na portáli `blog.matfyz.sk` za účelom zlepšenia manažmentu tagov. Náš prístup predpokladá aplikáciu sémantických technológií a metód učenia sa ontológií z textu za účelom rekurentnej konštrukcie ontológie nad textovým obsahom portálu. Koncepty v takto získanej ontológii úzko súvisia s tagmi, ktoré zadávajú autori príspevkov. Ontológiu potom využijeme pri sugescii

tagov v procese tagovania novo pridaného príspevku autorom. Očakávame nasledovný prínos pre náš portál:

- Zvýšenie relevancie tagov priradených používateľmi, vďaka automatickej sugescii tagov.
- Zlepšenie tematického prístupu k píspevkom, vďaka zvýšenej relevancii tagovania, ale aj vďaka novým podporným nástrojom na tematický prístup, ktoré nám práve ontológia umožní, napr. vylepšený tag-browser, odzrkadľujúci ontologické relácie medzi tagmi a pod.
- Objavovanie podobných príspevkov, vďaka ontologickým reláciám medzi tagmi

ľudských používateľov portálu, predovšetkým v procese selekcie tagov (t.j. extrakcie konceptov), ale tiež portáloví editori budú môcť dodatočne robiť zásahy do získanej ontológie. Zaujíma nás aký vplyv to bude mať na celkový výkon ontologického učenia, dúfame v jeho zlepšenie. Záverom je nutné poznamenať, že náš prístup je v súčasnosti vo fáze návrhu a je potrebné ho preveriť implementáciou. Na tejto implementácii v súčasnosti pracujeme.

Poďakovanie. Práca bola podporená projektom VEGA č. 1/0173/03 Ministerstva Školstva SR a Slovenskej Akadémie Vied. Naše poďakovanie patrí študentom, ktorí sa spolupodieľali na projekte portálu blog.matfyz.sk, predovšetkým Antonovi Kohutovičovi a Michalovi Novomeskému. Tiež ďakujeme Michaele Danišovej za jazykové úpravy.

Literatúra

[1] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini,*Ontology learning from text: An overview*, Ontology learning from text: Methods, evaluation and applications, IOS Press, 2005.

[2] Philipp Cimiano, Andreas Hotho, and Steffen Staab, *Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text*, Procs. of ECAI 2004, IOS Press, 2004, pp. 435-439.

[3] Philipp Cimiano, Aleksander Pivk, Lars Schmidt-Thieme, and Steffen Staab, *Learning taxonomic relations from heterogeneous sources of evidence*, Ontology learning from text: Methods, evaluation and applications, IOS Press, 2005.

[4] Ido Dagan, Oren Glickman, and Bernardo Magnini, *The pascal recognising textual entailment challenge*, Procs. of MLCW 2005, LNAI Volume 3944, Springer-Verlag, 2006, pp. 177-190.

[5] Zellig Harris, *Mathematical structures of language*, John Wiley and Sons,1968.

[6] Marti A. Hearst, *Automatic acquisition of hyponyms from large text corpora*}, Procs. of 14th International Conference on Computational Linguistics, 1992, pp. 539-545.

[7] Jason J. Jung and Jérôme Euzenat, *Towards semantic social networks*, The Semantic Web: Research and Applications, 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3-7, 2007, Proceedings, LNCS, vol. 4519, Springer, 2007, pp. 267-280.

[8] David R. Karger and Dennis Quan, *What would it mean to blog on the semantic web?*, Journal of Web Semantics **3** (2005).

[9] James B. MacQueen, *Some methods for classification and analysis of multivariate observations*, Procs. of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967.

[10] Fabrizio Sebastiani, *Machine learning in automated text categorization*, ACM Comuting Surveys **34** (2002), no. 1.

[11] Paola Velardi, Roberto Navigli, Alessandro Cuchiarelli, and Francesca Neri, *Evaluation of ontolearn, a methodology for automatic population of domain ontologies*, Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.