# Volba vlastností s využitím Linked Data

Martin DOSTAL[1], Karel JEŽEK[2]

[1]*Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni*
*Univerzitní 22, 309 02 Plzeň*
madostal@kiv.zcu.cz
[2]*Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni*
*Univerzitní 22, 309 02 Plzeň*
Jezek_ka@kiv.zcu.cz

**Abstrakt.** V tomto článku bychom chtěli představit metodu pro automatickou volbu vlastností s využitím Linked Data. Pojmem Linked Data jsou označovány techniky umožňující prezentování a publikování dat ve strojově čitelné podobě. My jsme těchto technik využili pro automatickou volbu vlastností, kterou lze použít např. pro shlukování. Tento přístup budeme demonstrovat na množině článků z oblasti IT, ke kterým byly automaticky přiřazeny štítky v podobě uzlů z Linked Data.

**Klíčová slova:** volba vlastností, shlukování, Linked Data.

## 1 Úvod

Problematika volby vlastností pro rozlišení skupin dokumentů je velmi aktuální téma a mezi jeho nejběžnější využití patří různé shlukovací metody. V současné době se shlukování využívá zejména v oblasti vyhledávání a prohlížení webu, kdy se může využívat buď pouze pro interní zpracování dokumentů, nebo ještě lépe pro prezentaci nalezených výsledků uživateli.

Mezi hlavní průkopníky sémantického vyhledávání s využitím shlukování patří vyhledávače Yippy [12] a Dogpile [10]. V prvním případě je již v praxi nasazeno velmi kvalitní vícejazyčné shlukování pro automatické zpracování výsledků z vyhledávacích portálů Google, Yahoo! a Bing. Tento způsob je velmi komfortní, neboť uživatel není nucen procházet stovky výsledků vyhledávání, ale pouze si omezí výsledky vyhledávání na shluk, který ho zaujal.

Pojem Linked Data zavedl Tim Berners Lee jako pojmenování technik publikace dat ve strojově čitelné podobě. Technicky je možné přidávat sémantické informace do existujících webových stránek s využitím formátu RDFa, nebo můžeme tyto informace zapisovat ve formě čistých RDF dokumentů. Nejvýznamnějším principem Linked Data je fakt, že každý pojem má jednoznačně přiřazené identifikační URL, díky kterému není třeba řešit desambiguaci pojmů. Každý pojem dále odkazuje na různé ontologie, na nadřazený pojem, na související pojmy a navíc i na pojmy podřízené. Díky tomu nám vzniká rozsáhlá stromová struktura, která čeká na naše využití.

Cílem tohoto článku je představení metody umožňující automatickou volbu vlastností s využitím Linked Data. V kapitole 2 si představíme základní principy Linked Data a existující metody zápisu hieratické struktury uzlů. Abychom mohli kvalitu volby vlastností vyhodnotit, aplikujeme jí na nejjednodušší a nejznámější shlukovací metodu K-průměrů. Základní principy a problémy shlukování si přiblížíme v kapitole 3.

Techniku volby vlastností s využitím Linked Data a její další možnosti rozšíření popíšeme v kapitole 4, kde se zaměříme zejména na problematiku analýzy štítků, se kterými budeme dále pracovat. Štítky jsou, v běžném smyslu, krátké textové řetězce neboli klíčová slova, používaná na Webu zejména k vyhledávání a třídění článků. Klasické členění článků do kategorií se v současné době jeví jako nepoužitelné a tak stále více webových stránek přistupuje k využití štítkování jako velmi efektivní metodě sdružování článků dle obsažených témat.

V kapitole 5 se zaměříme na problematiku vyhodnocení této metody. Vyhodnocení technik volby vlastností lze provádět buď přímo s využitím porovnání automatického seznamu vlastností s ručně stanoveným seznamem, nebo dle vlivu volby vlastností na nějaký algoritmus. V našem případě budeme vyhodnocení techniky volby vlastností provádět s využitím shlukovacího algoritmu K-průměrů.

V kapitole 6 si představíme cíl dalšího výzkumu zejména v oblasti vytvoření algoritmu pro shlukování přímo s využitím Linked Data a další možnosti pro vyhodnocení výsledků shlukování.

## 2 Linked Data

Nyní si stručně vysvětlíme, co to jsou Linked Data a jejich základní principy. Nejdříve bychom však měli začít pojmem Sémantický Web. Tim-Berners Lee definuje Sémantický Web [3] jako přístup umožňující vyjádření informací ve strojově čitelné podobě. Jeho hlavní myšlenka je provázání dat s využitím odkazů, což umožní lidem i strojům procházet související informace (Linked Data). T. B. Lee formálně definoval 4 základní pravidla Linked Data:

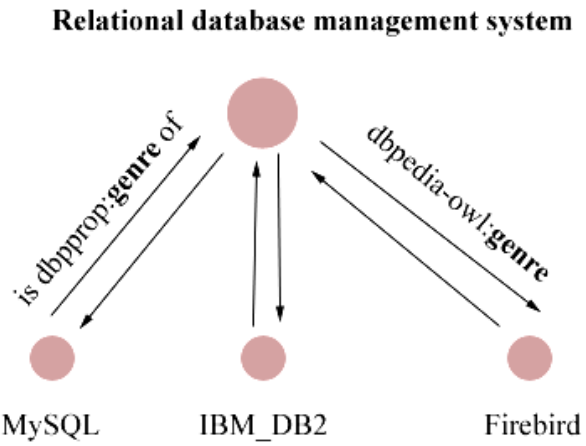
1. Používat URI jako identifikátory objektů.
2. Používat http URI, aby si lidé mohli tyto objekty prohlédnout.
3. Pokud se někdo podívá na URI, je třeba poskytnout užitečné informace s využitím standardů (RDF*, SPARQL).
4. Používat odkazy na další URI, což umožní vyhledávat související objekty.

První a druhé pravidlo vyžaduje používání URI jako identifikátorů. Jinak řečeno se jedná o jednoznačná jména, nikoliv adresy, jak by mohlo být mylně interpretováno. Třetí pravidlo vyžaduje popis objektu s využitím standardů v případě přístupu na toto URI. Člověk i stroj by měl získat základní informace o objektu, jeho formálním popisu a navíc i odkazy na další objekty, což je základní myšlenkou Linked Data.

Prakticky lze Linked Data získat např. z DBpedie [7], která obsahuje strojově čitelná data nashromážděná většinou z anglické Wikipedie. Tato data lze relativně jednoduše stáhnout a uložit do jedné z relačních databází. Záleží pouze na typech informací, které je třeba automaticky zpracovávat.

Odkazy na další uzly jsou realizovány s využitím ontologií, které určují typ vazby. V našem případě nás nejvíce zajímají tyto typy relací:

- **vztah potomek – rodič** – v Linked Data existuje odkaz většinou oběma směry. Lze využít relací: „*dbpprop:genre of*“ určující rodiče viz obr.1, „*skos:broader*“ a „*dcterms:subject*“ stanovující obecnější pojem. Možných relací je velmi mnoho a je tedy třeba používat ty, které máme k dispozici.
- **synonyma** – nejvhodnější je relace „*owl:sameAs*“ určující skutečná synonyma, případně však lze využít i relaci „*skos:related*“ označující související pojmy.




Obr. 1. Schéma hierarchické vazby mezi uzly v Linked Data [8]

## 3 Shlukování

Cílem algoritmu shlukování je automatické rozdělení množiny dokumentů na menší množiny podobných dokumentů. Množiny podobných dokumentů nazýváme *shluky*. V případě všech shlukovacích algoritmů lze definovat následující vstup a výstup:

**Vstup** - Počet požadovaných shluků *K* a množina dokumentů s jednoznačným označením. Jednoznačná identifikace dokumentu může být realizována s využitím ID, nebo prostým číslem dokumentu. Každý dokument se skládá z množiny slov ze slovníku *W*.

**Výstup** - Výstupem algoritmu je množina dokumentů přiřazených do shluků. Každý shluk je definován výčtem identifikátorů dokumentů, které do něj náleží.

V případě shlukování je možné zvolit jeden ze dvou základních přístupů:

- *Tvrdé shlukování* – každý dokument je přiřazen právě do jednoho shluku.
- *Měkké shlukování* – každý dokument může být přiřazen do více shluků.

Zatímco většina shlukovacích algoritmů využívá raději měkké shlukování než tvrdé, v našem případě se zaměříme právě na tvrdé shlukování, kde je možné provést jednodušší měření přesnosti a úplnosti dané metody.

Dle úrovně zařazení můžeme shlukování dále dělit na:

- *Jednoúrovňové shlukování* – všechny shluky jsou na stejné úrovni, součástí vstupu je většinou požadované množství shluků.
- *Hierarchické shlukování* – shluky se dále spojují a vytváří hierarchickou strukturu zcela automaticky, nebo dle vstupního požadavku na počet shluků, do kterých je třeba dokumenty rozdělit. Ve většině případů však není nutné zadávat požadované množství shluků, neboť algoritmy jsou schopné zvolit počet shluků samostatně.

V našem případě se opět z důvodu rychlejšího vyhodnocení budeme věnovat jednoúrovňovému shlukování jako nejjednoduššímu zástupci shlukování, které lze snáze vyhodnotit. V případě vyhodnocení hierarchického shlukování nastává zásadní problém v podobě rozhodnutí, kdy je shlukování považováno za správné a kdy už nikoliv. Zejména v případě umístění do velmi podobného shluku je třeba rozhodovat nikoliv o situaci určující

správné, nebo špatné zařazení, ale o míře správného zařazení. Tento fakt je však často zanedbáván, neboť neexistuje jeho efektivní a jednoznačné řešení [4].

Principem všech shlukovacích metod je tzv. shlukovací hypotéza, která tvrdí, že všechny dokumenty v rámci shluku se chovají podobně a mají podobné vlastnosti s ohledem na požadovanou informaci. Touto definicí je řečeno, že obsahují největší množství shodných termů nebo témat.

## 4 Volba vlastností s využitím Linked Data

V současné době existuje množství různých metod umožňující volbu vlastností [6]. Často se využívá korelačních koeficientů [6], vzájemné informace [1] nebo SVM [2].

Nyní si představíme techniku volby vlastností s využitím Linked Data. Algoritmus lze neformálně popsat následovně:

1. **Vstup** - Mějme množinu článků *D* očíslovaných *1,…, N* a množinu uzlů *L* z Linked Data, identifikovaných s využitím URI, tematicky pokrývající články. Jako alternativu je samozřejmě možné použít všechny uzly z Linked Data např. z DBPedie pokrývající většinu oblastí zájmu. Pro oblast IT se v současné době jedná řádově o 20 000 uzlů.
2. **Přiřazení štítků** - Článkům automaticky přiřadíme uzly z Linked Data a nazveme je štítky. Přiřazení štítků lze realizovat s využitím fulltextového vyhledávání názvů uzlů v kombinaci s obecně známými metodami jako je tokenizace, lemmatizace, odstranění stop-slov apod. Štítek je článku přiřazen v případě, kdy je tento pojem v textu článku nalezen aspoň 1x. Každý štítek může obsahovat i údaj určující jeho váhu vůči danému dokumentu. V nejjednodušším případě jí lze určit jako počet výskytů pojmu v textu článku. Cílem tohoto kroku je hrubé určení témat obsažených v rámci článků.
3. **Analýza štítků** – Štítky (uzly z Linked Data) obsahují množství zajímavých informací, kterých je možné využít k další analýze témat obsažených v článku. Cílem této analýzy je určení optimálních vlastností pro shlukování, kdy štítky budeme přímo považovat za jednotlivé vlastnosti. Tuto analýzu je třeba provádět na dvou různých úrovních:
    - *Lokální analýza* – v tomto případě budeme analyzovat množinu štítků u jednoho článku bez ohledu na štítky přiřazené k ostatním článkům.
    - *Globální analýza* – manipulace se štítky dle jejich rozšířenosti mezi články.
4. **Výstup** – Výsledkem je množina všech štítků obsažených ve všech článcích, které nejlépe definují články s ohledem na ostatní. Tyto štítky jsou přímo použitelné jako vlastnosti pro shlukování.

Dále si vysvětlíme hlavní principy analýzy štítků z lokálního i globálního hlediska. V případě lokální analýzy štítků zkoumáme štítky přiřazené ke článku bez ohledu na jejich četnost v rámci všech analyzovaných článků. Řešíme tedy pouze optimální volbu štítků, které by nejlépe popsaly témata ve článku obsažená. Základní techniky jsou následující:

- **záměna štítku jeho rodičem** – v tomto případě je příliš specifický štítek nahrazen obecnějším pojmem. Tato záměna je vhodná zejména v případě, kdy je ke článku přiřazeno více štítků se stejným rodičem. Touto záměnou dochází k výrazné redukci počtu přiřazených štítků. Místo několika méně významných štítků je tak přiřazen nový štítek s velkým významem pro daný článek.

V případě globální analýzy zkoumáme všechny štítky přiřazené ke všem článkům. V průběhu analýzy je samozřejmě možné použít základní techniky z lokální analýzy, avšak je třeba být opatrný, abychom vlastnosti příliš nezobecnili. Ve většině případů je vhodné využít vážených vlastností, které určí významnost zejména nově přiřazených štítků. Nyní si stručně představíme jednotlivé možnosti a upozorníme na nejvýznamnější problémy:

- **náhrada synonym** – v případě synonym lze zvolit verzi, která se v rámci článků nejčastěji vyskytuje a všechna ostatní synonyma touto verzí nahradit.
- **záměna štítku jeho rodičem a naopak** – v tomto případě je vhodné zjistit, zda by některé štítky nebylo vhodné sloučit, nebo naopak rozdělit, abychom dosáhli rozumné úrovně popisu. V první fázi je vhodné provést záměnu za rodiče a určit, zda výsledná množina článků sdílející tuto vlastnost nevytvoří shluk blížící se očekávané velikosti. Pokud ne, je vhodné pokusit se provést záměnu za potomka a otestovat, zda dojde k vytvoření shluku v případě kombinace několika štítků. V tomto případě však musíme dát pozor, aby se potomek, nebo některý z jeho následovníků v textu skutečně vyskytoval a nebyl tak přiřazen nesouvisející štítek.
- **odstranění jedinečných štítků** – velmi často dochází k přiřazení štítků, které se u daného článku vyskytují unikátně, a u žádného jiného článku je již nenajdeme. Pokud se nepodařilo aplikovat žádnou jinou metodu, je vhodné tyto štítky odstranit, neboť pro shlukování nemají žádný význam. Prakticky se může jednat o různá vzdálená témata, kterými se článek zabývá, ale které již všechny ostatní články ignorují, nebo nejsou dostatečně popsána v rámci Linked Data.

## 5 Vyhodnocení shlukování

V předchozí kapitole jsme představili techniku volby vlastností a nyní se podíváme na problematiku jejího vyhodnocení. Existují dvě základní metody vyhodnocení volby vlastností:

- **Přímé porovnání dvou sad vlastností** - porovnání sady automaticky zvolených vlastností s ručně zvolenými vlastnostmi. Jinak řečeno se snažíme automaticky zvolit stejné štítky charakterizující články, které by zvolil člověk.
- **Porovnání výsledků algoritmu –** porovnáme výstupy shlukovacího algoritmu, jehož vstupem budou jednou automaticky zvolené vlastnosti a podruhé ručně zvolené vlastnosti.

V případě přímého porovnání dvou sad vlastností nelze toto vyhodnocení považovat za správné a věrohodné, což je způsobeno zejména subjektivním výběrem štítků v případě manuálního výběru. Člověk může některé významné štítky vynechat a naopak jiné přidat. Samotné porovnání těchto dvou množin je sice velmi jednoduché, ale zároveň velmi nepřesné. Lze testovat pouze výskyt vs. absenci daného pojmu a nikoliv již blízkost automaticky a manuálně zvolené vlastnosti.

V případě porovnání výsledků algoritmu dojde k aplikaci dané sady vlastností a lze zkoumat její vliv na výsledná data. Nezáleží nám tak na konkrétních vlastnostech, ale spíše na tom, jaký mají vliv na stejný algoritmus realizující např. shlukování.

My jsme si zvolili druhou možnost, neboť ukazuje, že manuální volba vlastností nemusí být vždy vhodná a nemusí dosahovat nejlepších možných výsledků. Kromě toho je ruční volba vlastností zatížena subjektivním přístupem, který má v případě shlukování výrazně menší vliv a výsledek je tak lépe porovnatelný.

Jako základní shlukovací metodu jsme použili algoritmus K-průměrů umožňující rozdělení dokumentů do zadaného počtu shluků. Budeme využívat 3 základní techniky volby vlastností:

1. *Manuální volba vlastností* – vlastnosti jsou zvoleny ručně takovým způsobem, aby maximálně odlišily jednotlivé vzorové množiny dokumentů. Volbu vlastností provádí ručně uživatel na základně znalosti obsahu článků a znalosti názvů manuálních shluků, dle kterých budeme shlukování vyhodnocovat. Uživatel je dostatečně seznámen s problematikou shlukování.
2. *Volba vlastností s využitím statistické metody* – využíváme nejjednodušší statistický přístup v podobě TFIDF. V tomto případě chceme demonstrovat, že použití statistické metody je závislé na počtu dokumentů, na kterých se provede natrénování TFIDF skóre. V případě menšího množství dokumentů vrácených jako výsledky vyhledávání se ukazuje tato metoda jako nevhodná a navíc značně výpočetně náročná. Abychom mohli tuto techniku použít museli jsme zvýšit počet trénovacích dokumentů na 500. V případě trénování nad 25 dokumenty bylo TFIDF zcela nepoužitelné.
3. *Volba vlastností s využitím Linked Data* – naše metoda volby vlastností s využitím štítků získaných z Linked Data.

Jako testovací sadu jsme použili 10 x 25 ručně anotovaných článků, které měl shlukovací algoritmus správně rozdělit do pěti různých shluků. Vyhodnocení shlukování bylo prováděno na základě porovnání s ručním rozdělením dokumentů do daných shluků. Z důvodu věrohodnosti bylo spuštění shlukovacího algoritmu provedeno desetkrát, neboť v rámci inicializace centroidů metody K-means je využita náhodná složka pro prvotní nastavení centroidů, která má vliv na kvalitu výsledného shlukování.

V následující tabulce č. 1 jsou znázorněny výsledky shlukování s využitím manuálně zvolených vlastností. Význam sloupců je následující:

- Přesnost – standardní výpočet přesnosti.
- Úplnost – standardní výpočet úplnosti.
- Shoda – počet článků, kdy došlo ke shodě při porovnání s ručně stanovenými shluky.
- Navíc – kolik článků obsahují automatické shluky navíc a které do stanovených shluků nepatří.
- Chybí – počet článků, které nebyly v rámci automatických shluků nalezeny, avšak patřit tam dle manuálních shluků měly.

*Tab. 1. Shlukování s využitím manuálních vlastností*

| iterace | přesnost | úplnost | shoda | navíc | chybí |
|---|---|---|---|---|---|
| **1** | 0.64 | 0.64 | 16 | 9 | 9 |
| **2** | 0.76 | 0.76 | 19 | 6 | 6 |
| **3** | 0.457 | 0.64 | 16 | 19 | 9 |
| **4** | 0.48 | 0.48 | 12 | 13 | 13 |
| **5** | 0.4 | 0.4 | 10 | 15 | 15 |
| **6** | 0.321 | 0.36 | 9 | 19 | 16 |
| **7** | 0.351 | 0.52 | 13 | 24 | 12 |

| | | | | | |
|---|---|---|---|---|---|
| **8** | 0.645 | 0.8 | 20 | 11 | 5 |
| **9** | 0.613 | 0.76 | 19 | 12 | 6 |
| **10** | 0.56 | 0.56 | 14 | 11 | 11 |
| celkem: | | | **148** | **139** | **102** |

V následující tabulce č. 2 jsou znázorněny výsledky shlukování s využitím statisticky zvolených vlastností. Význam sloupců je shodný s předchozí tabulkou č. 1.

*Tab. 2. Shlukování s využitím vlastností dle TFIDF*

| iterace | přesnost | úplnost | shoda | navíc | chybí |
|---|---|---|---|---|---|
| **1** | 0.4 | 0.4 | 10 | 15 | 15 |
| **2** | 0.52 | 0.52 | 13 | 12 | 12 |
| **3** | 0.64 | 0.64 | 16 | 9 | 9 |
| **4** | 0.32 | 0.32 | 8 | 17 | 17 |
| **5** | 0.48 | 0.48 | 12 | 13 | 13 |
| **6** | 0.56 | 0.56 | 14 | 11 | 11 |
| **7** | 0.5 | 0.6 | 15 | 15 | 10 |
| **8** | 0.423 | 0.44 | 11 | 15 | 14 |
| **9** | 0.4 | 0.4 | 10 | 15 | 15 |
| **10** | 0.464 | 0.52 | 13 | 15 | 12 |
| celkem: | | | **122** | **137** | **128** |

V následující tabulce č. 3 jsou znázorněny výsledky shlukování s využitím vlastností získanými na základě našeho algoritmu využívající informace z Linked Data. Význam sloupců je shodný s předchozími tabulkami.

*Tab. 3. Shlukování s vlastnostmi získanými z Linked Data*

| iterace | přesnost | úplnost | shoda | navíc | chybí |
|---|---|---|---|---|---|
| **1** | 0.375 | 0.36 | 9 | 15 | 16 |
| **2** | 0.4 | 0.48 | 12 | 18 | 13 |
| **3** | 0.583 | 0.56 | 14 | 10 | 11 |
| **4** | 0.5 | 0.48 | 12 | 12 | 13 |
| **5** | 0.542 | 0.52 | 13 | 11 | 12 |
| **6** | 0.375 | 0.6 | 15 | 25 | 10 |
| **7** | 0.436 | 0.68 | 17 | 22 | 8 |
| **8** | 0.583 | 0.56 | 14 | 10 | 11 |

| | | | | | |
|---|---|---|---|---|---|
| **9** | 0.469 | 0.6 | 15 | 17 | 10 |
| **10** | 0.472 | 0.68 | 17 | 19 | 8 |
| celkem: | | | **138** | **159** | **112** |

Nyní si porovnáme celkově dosažené výsledky u všech třech typů shlukování. Z následující tabulky vyplývá, že nejvhodnější volba vlastností je samozřejmě ruční, což se však dalo předpokládat. Naše metoda volby vlastností je však jen o 8% horší, v případě porovnání F-míry. Statistická volba vlastností s využitím TFIDF dosáhla také velmi dobrých výsledků, avšak pouze díky použití dvou významných modifikací:

- TFIDF bylo trénováno na 20x větší množině článků, neboť v případě trénování na stejně velké množině článků se tato technika projevila jako nepoužitelná.
- Výsledné štítky zvolené dle TFIDF byly upraveny tím způsobem, že bylo vynecháno 5% štítků s nejvyšším TFIDF skóre, neboť byly příliš unikátní a v rámci shlukování neměly význam. Dále byl počet vlastností získaných s využitím TFIDF omezen na nejlepších 30% a zároveň na maximální počet vlastností ve výši 40 kusů.

*Tab. 4. Porovnání výsledků shlukování*

| volba vlastností: | **manuální** | **tfidf** | **Linked Data** |
|---|---|---|---|
| přesnost: | 0,516 | 0,471 | 0,465 |
| úplnost: | 0,592 | 0,488 | 0,552 |
| F-míra: | 0,551 | 0,479 | 0,505 |

## 6 Budoucí práce

Cílem další práce bude zejména vyhodnocení této techniky volby vlastností s využitím shlukování na větší kolekci dokumentů. K tomuto účelu je však třeba použít kolekci dokumentů nejlépe z oblasti IT obsahující informace o zařazení článků do shluků, kterou jsme doposud neměli k dispozici. Na základě inspirace článkem [2] bychom chtěli tuto techniku ověřit na projektu Open Directory [11], který obsahuje hierarchickou strukturu kategorií včetně zařazených článků. Dle [2] je navíc možné ze služby Delicious [9] získat k těmto článkům štítky v podobě krátkých textových řetězců, což by mohla být další zajímavá a využitelná informace. Z důvodu porovnatelnosti s [2] bychom chtěli použít stejnou transformaci hierarchické struktury kategorií z ODP do plošných a tvrdých kategorií.

Naše další práce se bude zabývat i metodami shlukování přímo s využitím struktury Linked Data. Nebudeme tak již používat žádné vlastnosti vstupující do existujícího algoritmu, ale pokusíme se navrhnout vlastní algoritmus, který přímo na základě informací z Linked Data navrhne optimální počet shluků a články rozdělí do skupin dle souvislostí mezi tématy. Prakticky nám tak do algoritmu vstoupí problematika analýzy velmi rozsáhlého grafu s výpočtem, nebo odhadem délky cesty mezi uzly.

Naše technika volby vlastností s využitím Linked Data bude dále používána a testována na vlastní kolekci obsahující 15 000 článků z oblasti IT typu Call for papers, které však nebylo možné v současné době využít z důvodu chybějících anotací. Tyto pozvánky

k účasti na konferenci byly nashromážděny z veřejně dostupných zdrojů, jako jsou např. emaily, diskusní fóra nebo přímo webové stránky konferencí.

## 7 Závěr

V tomto článku jsme představili techniku umožňující automatickou volbu vlastností s využitím strojově čitelných informací z Linked Data. Popsali jsme si vlastní algoritmus výběru vlastností a přiblížili hlavní problémy vyhodnocení volby vlastností.

Jako vhodný přístup k vyhodnocení volby vlastností se jeví algoritmus shlukování, který tyto vlastnosti využívá. Nehodnotíme tak tedy přímo získané vlastnosti, ale jejich význam pro další zpracování, což je mnohem více vypovídající hodnotou.

Tato technika volby vlastností s využitím Linked Data dosáhla v tomto případě jen o 8% horších výsledků při porovnání F-míry vzhledem k manuální volbě vlastností.

## Poděkování

Tato práce byla částečně podpořena z projektů: CZ.1.05/1.1.00/02.0090 *„NTIS“* a GAČR P103/11/1489 *„Metody pro tvorbu a ověřování komponentových systémů ze specifikací v přirozeném jazyce“*.

## Literatura

1. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. In: *Proceedings of Artificial Intelligence*, 1997.
2. Bradley, P. S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: *Proceedings of ICML*, 1998.
3. Lee, T. B. Linked Data – Design Issues, W3C. http://www.w3.org/DesignIssues/LinkedData.html.
4. Fischer, B., Buhmann, J. H.: Data Resampling for Path Based Clustering. In: *Pattern Recognition*, Springer Berlin/Heidelberg, 2002, 206 – 214.
5. Ramage, D., Heymann, P., Manning, D. Ch, Garcia-Molina, H.: Clustering the Tagged Web. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, R.E. Miller and J.W. Thatcher (Eds.), ACM New York, NY, USA (2012), 54-63.
6. Wolf, L., Shashua, A.: Feature Selection for Unsupervised and Supervised Inference: the Emergence of Sparsity in a Weighted-Based Approach. In: *Proceedings of Ninth IEEE International Conference on Computer Vision*, 2003, 378- 384 vol.1.
7. DBpedia. http://dbpedia.org.
8. DBpedia – Relational database management system. http://dbpedia.org/page/Relational_database_management_system
9. Delicious. http://delicious.com.
10. Dogpile search engine. http://dogpile.com.
11. Open Directory Project. http://dmoz.org.
12. Yippy search engine. http://yippy.com.

*Volba vlastností s využitím Linked Data*

**Annotation:**

*Feature selection with Linked Data application*

In this paper we would like to present our approach to feature selection with application of Linked Data. Linked Data can be defined as a set of rules and techniques for connections between Semantic Web resources. The importance of Linked Data is in links, so that a person or machine can explore the web of data. We used these techniques and resources for automatic feature selection and applied it for clustering method. We will demonstrate this approach with data collection related to IT area. These articles were automatically expanded by tags defined as resources from Linked Data. We will explain the transformation of these tags into features that can be used for clustering.