Klasifikace Suffix Tree frázemi - srovnání s metodou Itemsets

Roman Tesař [1], Karel Ježek [1]

[1]Katedra Informatiky a výpočetní techniky, Západočeská Univerzita v Plzni,
Univezitní 8, 306 14, Plzeň
{romant, jezek_ka}@kiv.zcu.cz

Abstrakt. V článku je prezentován postup klasifikace pomocí Suffix Tree (ST) frází. Popsán je způsob jejich získání, ohodnocení a použití ke klasifikaci textů. Konzultovány jsou výhody a nevýhody tohoto postupu, které jsou průběžně srovnávány s metodou Itemsets, ze které princip klasifikace Suffix Tree frázemi vychází. Popsán je také způsob nastavení prahové hodnoty pro zařazení dokumentu do zvolených tématických tříd. Prostor je věnován i porovnání vlivu průměrné délky dokumentu na celkovou úspěšnost klasifikace a srovnáván je také vliv itemsetů a ST-frází vyšších řádů u obou metod. V závěru samozřejmě nechybí srovnání dosažených výsledků klasifikace s dalšími rozšířenými metodami klasifikace textů.

Klíčová slova: klasifikace, kolekce dokumentů, Itemsets, ST-fráze, Suffix Tree, ohodnocení dokumentu, nastavení prahu

1 Úvod

V elektronické podobě se v současné době vydává stále více článků, časopisů i knih. Je v nich obsaženo obrovské množství informací, avšak ne všechny jsou pro každého uživatele zajímavé. Aby bylo možné se v této záplavě orientovat, je potřeba stanovit, jaký typ informací a z jaké oblasti se v jednotlivých článcích nachází. Existují již komerční instituce, které se zabývají vyhledáváním pouze relevantních informací pro své klienty. Jsou však pryč doby, kdy na tuto činnost byla vyhrazena skupina lidí pročítajících elektronické texty z různých zdrojů. Použití aplikací implementujících různé metody automatické klasifikace textu je vzhledem k stále většímu objemu dat nezbytností.

Klasifikátory využívající princip strojového učení ovšem potřebují ke svému natrénování a k následnému zařazení klasifikovaného článku množinu textů zastupující jednak jednotlivé oblasti, které jsou pro daného uživatele relevantní a jednak oblast, která je pro uživatele nezajímavá. Pokrýt ovšem opravdu všechny existující tématické okruhy, které nejsou pro daného uživatele zajímavé, není vzhledem k jejich obsahové šířce možné.

V tomto článku se věnujeme metodě využívající ke klasifikaci textu ohodnocených Suffix Tree frází, která tento nedostatek odstraňuje. Její princip a dosažené výsledky průběžně konzultujeme s metodou Itemsets, která je svým principem velmi podobná a která dovoluje několik modifikací.

2 Vlastnosti metody Itemsets

Při návrhu a konstrukci klasifikátoru využívajícího Suffix Tree fráze jsme vycházeli z principu metody Itemsets popsané v [3], [4] a [5], která úspěšně klasifikuje zejména krátké textové dokumenty. Její nevýhodu bohužel představuje především fáze trénování, kdy jsou pomocí Apriori algoritmu (viz [5]) hledány časté množiny slov (itemsety). Složitost tohoto algoritmu je pro 1-itemsety lineární. Pro nejhorší případ je s rostoucím K při hledání K-itemsetů vyšších řádů kombinatorická, a to jak složitost časová, tak i složitost paměťová. Výhodu ale představuje rychlá fáze klasifikace této metody spočívající v přičítání ohodnocení předem natrénovaných K-itemsetů, jejichž všechny jednotlivé termy (slova) se v dokumentu vyskytují současně.

Zajímavá je skutečnost, že použití K-itemsetů, kde K>1, nepřineslo žádné zlepšení úspěšnosti klasifikace (viz kapitola 4). V důsledku potom použití 1-itemsetů, tedy jednotlivých slov, přináší této metodě nejvyšší úspěšnost. Dle našeho názoru je to dáno faktem, že jednotlivá slova K-itemsetu nalezeného Apriori algoritmem spolu netvoří souvislou frázi, ale jen množinu slov často se společně vyskytujících v trénovacích dokumentech. To je bezpochyby použitelné pro krátké dokumenty obsahujících řádově desítky slov, které jsou pokud možno tématicky dostatečně vzdálené. Pro delší dokumenty obsahující větší počet slov již úspěšnost klasifikace klesá v důsledku zvětšující se pravděpodobnosti nalezení většího počtu K-itemsetů společných několika klasifikačním třídám.

Stejný problém nastává, pokud budeme chtít klasifikovat do více tříd dokumenty tématicky velmi podobné. Opět budou Apriori algoritmem nalezeny K-itemsety obsahující stejná nebo podobná slova pro více tříd, což může vést ke snížení úspěšnosti klasifikace. V případě, že bychom chtěli klasifikovat dokumenty jen do dvou tříd (například závadné/nezávadné), je navíc nezbytné mít k dispozici kolekce dokumentů pro obě třídy, abychom mohli klasifikátor založený na metodě Itemsets natrénovat.

Na základě těchto poznatků jsme vytvořili klasifikátor, který uvedené nedostatky umožňuje alespoň částečně odstranit. K jeho konstrukci jsme využili fráze získané algoritmem Suffix Tree.

2.1 Suffix Tree fráze versus Itemsety

Oproti K-itemsetům nejsou Suffix Tree fráze (dále jen ST-fráze) množiny slov současně se vyskytujících s určitou četností v trénovacích dokumentech (dokumentech zařazených do třídy). Jedná se o posloupnost po sobě následujících slov, které se v množině trénovacích dokumentů vyskytují se zadanou četností. Intuitivně by tedy ST-fráze měly lépe vystihovat tématické okruhy představující jednotlivé klasifikační třídy a to i v případě, že si budou velmi podobné. Vycházíme přitom z předpokladu, že tématicky blízké dokumenty budou obsahovat stejná nebo podobná slova. Na následujícím jednoduchém přikladu jsme si ověřili, že metoda Itemsets by většinu takovýchto dokumentů klasifikovala do stejné třídy.

Tab. 2.1 Několik vybraných natrénovaných 1, 2 a 3-itemsetů pro závadnou třídu

1-Itemsety	ohodnocení
free	93,25
adult	88,25
hardcore	83,06
teen	79,75
picture	70,81
xxx	67,43
gallery	66,25

2-itemsety	ohodnocení
you hardcore	73,18
amateur free	77,25
girl teen	72,93
free hot	67,25
xxx adult	60,53
gallery picture	56,34
site free	48,12

3-itemsety	ohodnocení
site hardcore free	70,32
adult girl teen	68,58
maiden babe your	59,15
free site male	53,80
woman school asia	47,65
guy great xxx	42,30
latina lady adult	34,24

K jejímu natrénování jsme použili 200 závadných internetových stránek sexuálního charakteru[2] a 2000 nezávadných stránek obsahujících různá témata. Jednotlivé třídy (závadné/nezávadné) charakterizovalo 60 nejčastějších K-itemsetů pro $K \leq 3$ vytvořených Apriori algoritmem. Testovali jsme nezávadné stránky, u kterých jsme očekávali chybnou klasifikaci na základě předpokladů vysvětlených v následujícím odstavci. K výpočtu kvalitativní váhy itemsetů jsme v průběhu veškerých testů použili stejný vztah, který se osvědčil a byl použit i v [5].

Ačkoli na adrese *www.bhs.org.uk/Horse_clip-art/icons.htm* žádný sexuálně orientovaný materiál nenajdeme, byla klasifikována metodou Itemsets jako závadná. Důvod je ten, že se v HTML kódu této stránky vyskytují termy "free, gallery, picture". Tyto termy se bohužel vyskytují mezi častými itemsety charakterizujícími závadnou třídu a protože zde již není mnoho textu, ve kterém by mohly být nalezeny další natrénované termy patřící do nezávadné třídy, je tato stránka chybně klasifikována. Této skutečnosti také napomáhá fakt, že váhu klasifikace nesou 1-itemsety jak již bylo zmíněno dříve. Nelze tedy účinně přenést váhu klasifikace na itemsety vyšších řádů nastavením nižších váhových koeficientů 1-itemsetům. Navíc i některé natrénované 2 a 3 itemsety, jak je patrné z Tab. 2.1, mohou být závádějící. Pokud se například v textu dokumentu objeví trojice termů "site hardcore free", nemusí se vůbec jednat o závadnou stránku, možná právě naopak. Příkladem jsou i následující stránky, u kterých dochází k podobným problémům:

http://spgm.sourceforge.net/
http://www.ag.ohio-state.edu/~vegnet
http://www.travelplan.com/gallery4.htm
http://www.hikyaku.com/gallery/gallery.html
a další…

Vznikají zde ovšem i jiné potíže. Ohodnocení přiřazené K-itemsetům závadné třídy je většinou velmi vysoké, protože stránky sexuálního charakteru jsou velmi úzce zaměřeny a obsahují často stejná slova (1-itemsety) a mnohdy i slovní spojení. To samozřejmě neplatí pro nezávadnou třídu, zejména pokud se snažíme jednotlivými dokumenty pokrýt co nejvíce tématických okruhů a oblastí lidské činnosti. V důsledku tedy, pokud se v závadné trénovací kolekci bude vyskytovat dostatečně často nezávadné

[2] Tento článek vznikl během projektu na filtraci závadných internetových stránek, proto kolekce takového typu

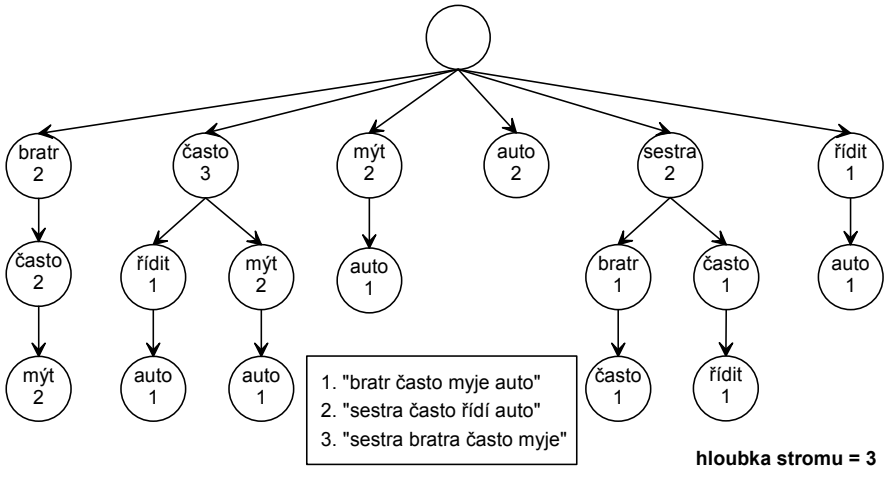
slovo, natrénuje se jako závadný itemset. Pokud se potom v klasifikovaném dokumentu takové slovo objeví, může toto slovo svým ohodnocením zapříčinit zařazení tohoto dokumentu do závadné třídy, protože i při současném výskytu několika natrénovaných nezávadných slov nemusí jejich ohodnocení stačit na zařazení dokumentu do správné třídy. Navíc ani pokud bychom měli k dispozici obrovskou kolekci nezávadných dat nemůžeme pokrýt tuto oblast zcela. Nebylo by ani možné použít Apriori algoritmus pro natrénování K-itemsetů vyšších řádů z důvodu jeho příliš velkých časových a paměťových nároků (viz [3] a [5]).

Z těchto poznatků jsme usoudili, že podobné problémy mohou nastat i při klasifikaci do více tříd. Jak bude uvedeno dále, ke klasifikaci pomocí ST-frází postačuje mít k dispozici jen data závadné třídy, není nutné vytvářet kolekci nezávadných dokumentů. Tento fakt přináší samozřejmě snížení celkové složitosti, protože postačuje vytvořit ST-fráze jen pro třídu zastupující úzký tématický okruh, který chceme rozpoznávat.

3 Suffix Tree klasifikátor

3.1 Fáze trénování

Při trénování jsou v kolekci vybraných dokumentů, které jsou tématicky zaměřeny na oblast, kterou budeme chtít v budoucnu rozpoznávat, vyhledávány opakované fráze. K jejich získání využíváme postupného vytváření Suffix Tree struktury zadané hloubky (princip a postup viz [1] a [2]). V jednotlivých uzlech ovšem uchováváme jen informaci o počtu výskytů dané fráze v trénovací kolekci (viz Obr. 3.1) a název termu (slova). Další údaje nejsou pro získání ohodnocených ST-frází nutné.



Obr. 3.1 Příklad struktury Sufix Tree vytvořené z trénovací kolekce tří dokumentů

Trénovací kolekci dokumentů, ze které chceme strukturu vytvořit, je vhodné nejprve lematizovat. Na kolekci, která nebyla lematizována, jsme v průběhu testování dosahovali vždy průměrně o 5% horší úspěšnosti klasifikace. Protože jsme testovali uspěšnost metody Sufix Tree a Itemsets jen na anglických textech (viz kapitola 4), použili jsme Porterův lemmatizátor (viz [7]) pro anglický jazyk, jehož programová realizace je dostupná z [9]. Slovník stopslov obsahoval 390 nejčastějších anglických slov převzatých z WAIS[3]. Ty je samozřejmě vhodné před počátkem trénování odstranit. Text jednotlivých dokumentů byl vždy zpracováván jako jeden dlouhý řetězec, na konce vět nebyl brán zřetel.

Z výsledné Sufix Tree struktury následně získáme jednotlivé fráze, jejichž maximální délka je rovna maximální hloubce stromu a na základě vzorce

$$F_{ohod} = \frac{F_{čet}}{N_{K,\max}} \cdot 100 \ [\%] \qquad (1)$$

určíme jejich ohodnocení. Jednotlivé symboly mají následující význam:

F_{ohod} = výsledné ohodnocení konkrétní fráze
$F_{čet}$ = četnost konkrétní fráze v trénovací kolekci
K = délka ST-fráze
$N_{K,max}$ = počet výskytů nejčastější fráze (délky K) v trénovací kolekci

Jak je ze vzorce (1) patrné, pro ST-fráze různých délek se mění hodnota ve jmenovateli sloužící jako normovací konstanta. Je to z toho důvodu, aby nebylo potlačeno

Tab. 3.1 Ohodnocené ST-fráze získané ze struktury na Obr. 3.1

ST-fráze	$F_{čet}$	F_{ohod} [%]	K (délka ST-fráze)	$N_{K,\max}$
často	3	100	*1*	*3*
bratr	2	67		
mýt	2	67		
auto	2	67		
sestra	2	67		
řídit	1	33		
bratr často	2	67	*2*	*2*
často myje	2	67		
často řídit	1	33		
mýt auto	1	33		
sestra bratr	1	33		
sestra často	1	33		
řídí auto	1	33		
bratr často mýt	2	67	*3*	*2*
často řídit auto	1	33		
často mýt auto	1	33		
sestra bratr často	1	33		
sestra často řídit	1	33		

[3] http://fog.bio.unipd.it/waishelp/stoplist.html

ohodnocení ST-frází větších délek. Použijeme-li totiž k normování jen nejčastěji se vyskytující frázi, bude se vždy jednat o frázi délky 1 a tím by došlo k znevýhodnění delších frází. Proto musíme normovat vždy hodnotou, která je rovna počtu výskytů nejčastější fráze délky odpovídající ohodnocované frázi.

Tabulka Tab. 3.1 obsahuje příklad již ohodnocených ST-frází délky 1,2 a 3 získaných ze Suffix Tree struktury zobrazené na Obr. 3.1, která byla vytvořena ze tří dokumentů uvedených na stejném obrázku.

3.2 Fáze klasifikace

Ve fázi klasifikace nejprve vybereme určitý počet ST-frází, které mají nejvyšší ohodnocení určené na základě vzorce (1). Vliv počtu a délky ST-frází na úspěšnost klasifikace diskutujeme v podkapitolách 4.1 a 4.2.

Testovaný dokument procházíme a hledáme výskyt jednotlivých ST-frází. Tento postup je obdobný jako u metody Itemsets, oproti které ovšem nehledáme současný výskyt jednotlivých termů K-itemsetu kdekoli v dokumentu. Jednotlivé termy ST-fráze délky větší než 1 se v dokumentu musí vyskytovat těsně za sebou ve stejném pořadí jako při svém natrénování. Při výskytu některé z ST-frází charakterizujících určitou třídu přičteme pro testovaný dokument této třídě celkovou váhu odpovídající ohodnocení nalezené ST-fráze. Po otestování výskytu všech vybraných ST-frází reprezentujících jednotlivé klasifikační třídy zařadíme testovaný dokument do třídy s nejvyšší vahou V_{max}. Dalším rozdílem je skutečnost, že metoda Itemsets uvažuje jen prostý výskyt jednotlivých itemsetů v dokumentu. Při klasifikaci Suffix Tree frázemi uvažujeme vícenásobný výskyt jednotlivých frází v testovaném dokumentu.

Samozřejmě můžeme chtít asociovat dokument s více třídami. V tom případě zařadíme klasifikovaný dokument do všech tříd, jejichž váha přesahuje zvolený práh z maximální dosažené váhy V_{max} (jedná se tedy o zvolené procento z V_{max}). V průběhu celého testování jsme dosahovali nejlepších výsledků s hodnotu prahu 75 % pro obě porovnávané metody.

4 Testování, dosažené výsledky

Pro porovnání vlastností obou metod jsme vytvořili dvě kolekce složené z krátkých a dlouhých dokumentů. Využili jsme k tomu dokumenty anglické kolekce Reuters Corpus Volume 1 z týdnů 1 až 5 zařazených pouze do kořenových tříd (Economics, Markets, Government/Social, Corporate/Industrial). Následující tabulka (Tab. 4) obsahuje charakteristiky těchto kolekcí.

Tab. 4 Charakteristiky použitých kolekcí

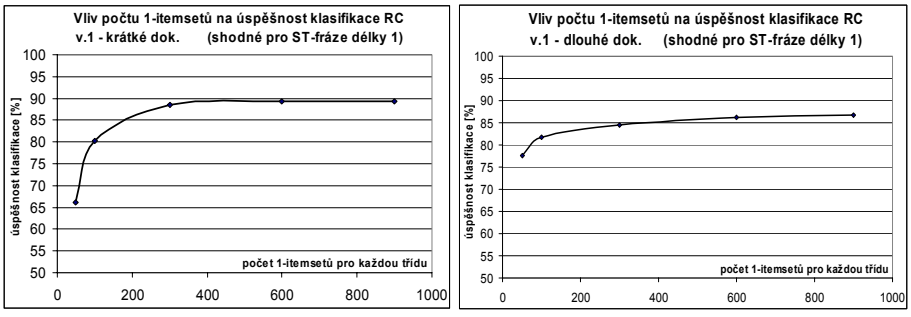
RC v.1 – KRÁTKÉ dokumenty		RC v.1 - DLOUHÉ dokumenty	
počet dokumentů celkem	9370	počet dokumentů celkem	10350
počet trénovacích dokumentů	2343	počet trénovacích dokumentů	2588
počet testovacích dokumentů	7028	počet testovacích dokumentů	7762
počet významových slov v dokumentech	**50 – 150 průměrně : 88**	počet významových slov v dokumentech	**300 – 3952 průměrně : 560**
počet všech slov	7244561	počet všech slov	10747726
počet významových slov	4306252	počet významových slov	6872532
počet různých významových slov	46653	počet různých významových slov	83900
počet nevýznamových slov	2938310	počet nevýznamových slov	3875194
průměrný počet tříd, do kterých je dokument klasifikován	1,5	průměrný počet tříd, do kterých je dokument klasifikován	1,8

4.1 Vliv počtu 1-itemsetů (ST-frází délky 1) na úspěšnost klasifikace

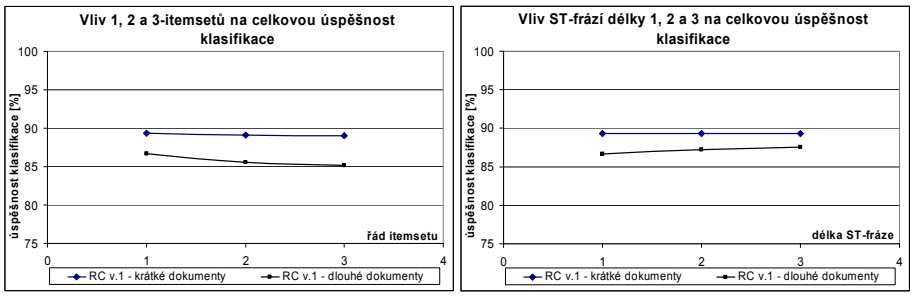
Při testování obou metod jsme se nejprve zaměřili na srovnání vlivu počtu 1-itemsetů na úspěšnost klasifikace. Míra celkové úspěšnosti byla v průběhu celého testování chápána jako průměr dosažených hodnot přesnosti a úplnosti. Protože natrénované 1-itemsety odpovídaly (včetně svého ohodnocení) natrénovaným ST-frázím délky 1 a protože princip fáze klasifikace je shodný, jsou výsledky z grafu 4.1 (kde je úspěšnost chápána jako průměr přesnosti a úplnosti) společné pro obě srovnávané metody.

Jak je z grafů patrné, u obou metod nastává pro krátké dokumenty od určitého počtu termů reprezentujících jednotlivé třídy "nasycení" a přidávání dalších termů již nemá větší vliv na úspěšnost klasifikace. Pro malý počet termů (50) samozřejmě docházelo k jevu, kdy v mnoha dokumentech nebyl nalezen žádný 1-itemset reprezentující některou z tříd. Z grafů je také zřejmé, že vhodným kompromisem mezi kvalitou klasifikace a složitostí procesu trénování bylo použití cca 400 termů.

Naopak pro kolekci delších dokumentů přidávání dalších termů reprezentujících jednotlivé třídy mělo smysl. S jejich vzrůstajícím počtem vzrůstala i úspěšnost klasifikace. Dochází zde však postupně k podobnému jevu jako v předchozím případě. Pro 800 termů a více se již projevuje vliv "nasycení". Počet 1-itemsetů (ST-frází délky 1) je tedy vhodné volit na základě předpokládané délky klasifikovaných dokumentů.



Graf 4.1 Porovnání vlivu počtu 1-itemsetů na celkovou úspěšnost klasifikace



Graf 4.2 Porovnání vlivu itemsetů a ST-frází vyšších řádů na úspěšnost klasifikace

Použití většího počtu by vedlo k zbytečnému zpomalení klasifikace dokumentů, použití nižšího počtu termů potom vede ke snížení celkové úspěšnosti klasifikace.

4.2 Vliv 1, 2 , 3-itemsetů a ST-frází délky 1, 2, 3

Dalším předmětem našeho testování byla klasifikace s použitím itemsetů a ST-frází vyšších řádů. Hodnoty z grafu 4.2 jen potvrzují co je již intuitivně zřejmé. U krátkých dokumentů nehrají delší ST-fráze významější roli. Zajímavé je však zjištění, že celková úspěšnost klasifikace u metody Itemsets je dána především 1-itemsety. Nemá tedy smysl generovat poměrně náročným Apriori algoritmem K-itemsety, kde K>1. Jak je z grafů patrné, úspěšnost se jejich použitím pro dlouhé dokumenty může zmenšovat. Je to pravděpodobně dáno skutečností, že s rostoucí délkou dokumentů se do procesu klasifikace zavlékají další obecné termy, které však pro svou obecnost nelze chápat jako charakteristické pro třídy, do kterých dokument skutečně patří. Větší délka dokumentů potom zvyšuje pravděpodobnost jejich případného současného výskytu a tím i možnost nesprávné klasifikace.

Pro krátké dokumenty již není vysvětlení mírného poklesu úspěšnosti klasifikace při použití itemsetů vyšších řádů tak jednoduché. Důvodem by mohl být stále příliš velký počet termů v testovacích dokumentech a skutečnost, že jednotlivé třídy jsou svým tématickým zaměřením poměrně široké. Natrénované 2 a 3-itemsety reprezentující jednotlivé třídy totiž měly výrazně menší váhu než itemsety získané v testu popsaném v podkapitole 2.1.

Tab. 4.2 Dosažené výsledky klasifikace pro jednotlivé metody

P = přesnost U = úplnost	**Naive Bayes**		**NBCI**		**TFIDF**		**Itemsets** (1,2,3)		**ST-fráze** (1,2,3)	
	P	U	P	U	P	U	P	U	P	U
RC v.1 – Krátké dokumenty	94,22	85,28	89,38	80,68	84,83	84,55	89,42	89,37	89,52	89,4
	89,75		***85,03***		***84,69***		***89,40***		***89,46***	

	Naive Bayes		**NBCI**		**TFIDF**		**Itemsets** (1,2,3)		**ST-fráze** (1,2,3)	
	P	U	P	U	P	U	P	U	P	U
RC v.1 – Dlouhé dokumenty	91,17	78,46	89,06	76,34	86,29	87,1	85,45	87,86	86,74	88,95
	84,82		***82,70***		***86,70***		***86,66***		***87,85***	

Naopak použitím delších ST-frází úspěšnost klasifikace roste, což se nejvíce projeví u delších dokumentů. Je to samozřejmě dáno jejich větší schopností rozlišit i velmi blízká klasifikační témata. Současně je oproti itemsetům výhodou menší pravděpodobnost jejich výskytu v dokumentech, které nepatří do třídy, kterou reprezentují.

4.3 Celkové výsledky klasifikace

V Tab. 4.2 jsou prezentovány konečné výsledky klasifikace na kolekci dlouhých a krátkých dokumentů. Hodnoty z této tabulky by neměly implikovat obecnou kvalitu jednotlivých klasifikačních metod, pro jiná vstupní data můžeme samozřejmě dostat jiné pořadí úspěšnosti prezentovaných metod.

Nejlepší hodnoty pro krátké dokumenty bylo dosaženo metodou Naive Bayes. Ta používá ke klasifikaci všech slov trénovací kolekce, což je v tomto případě nejlepší přístup. Nedochází zde totiž k zbytečné ztrátě informace vlivem hledání určitých častých položek, jako je tomu u ostatních metod. Již u krátkých dokumentů je patrný malý náskok v úspěšnosti klasifikace, který získávají ST-fráze oproti Itemsetům. Tento náskok se podle očekávání nejvíce projevil u kolekce delších dokumentů, kde již rozdíl mezi metodou Itemsets a ST-frázemi činil 1,2%. Delší dokumenty také podle očekávání lépe vyhovují metodě TFIDF, které umožnuje přítomnost většího počtu termů vytvořit relevantnější vektory reprezentující jednotlivé dokumenty a ve fázi klasifikace potom dochází k méně chybám. Popis méně známé metody NBCI je možné nalézt v [6].

Jak je z Tab. 4.2 patrné, obecně lepších výsledků jednotlivé metody dosáhly při klasifikaci kolekce krátkých dokumentů. Delší dokumenty pravděpodobně obsahují větší počet jednotlivých termů společných více třídám, což ztěžuje klasifikaci.

5 Závěr

Na testovacích kolekcích jsme ověřili, že výsledky klasifikace ST-frázemi jsou velmi dobré i v porovnání s jinými známými metodami. Použití delších ST-frází vede k lepším celkovým výsledkům než u metody Itemsets, což je nejvíce patrné u delších dokumentů. Použití 2 a 3-itemsetů nepřineslo žádné zlepšení. Naopak se celková úspěšnost klasifikace zhoršila.

Fáze trénování při klasifikaci ST-frázemi je algoritmicky méně náročná než u metody Itemsets. Složitost Apriori algoritmu totiž při hledání K-itemsetů výrazně roste se zvyšující se hodnotou K (viz [4]), zatímco vytváření Suffix Tree struktury má stále lineární složitost (viz [8]). Fáze klasifikace je u obou metod algoritmicky i časově srovnatelná.

Předmětem našeho dalšího zkoumání bude porovnání složitosti různých algoritmů pro získání ST-frází, testování možností dalšího zvyšování efektivity tohoto způsobu klasifikace a možnost kombinace ST-frází s dalším klasifikátorem založeným především na metodě Naive Bayes.

Reference

1. Grolmus P., Hynek J., Ježek K. User Profile Identification Based on Text Mining. Proc. of *6th Int. Conf. ISIM'03*, pp. 109-118, MARQ Ostrava, ISBN 80-85988-84-4
2. Grolmus P., Hynek J., Ježek K. Vyhledávání častých frází pro generování uživatelských profilů (in czech), *ITAT 2003,* pp.21-29, ISBN 80-7097-564-4
3. Hynek J., Jezek K. Document Classification Using Itemsets, *Proc.34th Int. Conf. MOSIS 2000, ISM 2000,* pp.97-102, ISBN 80-85988-45-3
4. Hynek J., Jezek K., Rohlik O. Short Document Categorization - Itemsets Method, *PKDD 4-th European Conference on Principles and Practice of Knowledge Discovery in Databases, Workshop Machine Learning and Textual Information Access.* Lyon, France, Sept.2000, pp.14-19
5. Hynek J., Ježek K.: Automatická klasifikace dokumentů do tříd metodou Itemsets, její modifikace a vyhodnocení (in czech), *Datakon 2001*, pp. 329-336, ISBN 80-227-1597-2
6. Kučera M., Ježek K., Hynek J. Text Categorization Using NBCI Method (in czech), *ZNALOSTI 2003*, Proc. pp.33-42, VŠB Technická univerzita Ostrava, ISBN80-248-0229-5
7. Porter, M.F., 1980, An algorithm for suffix stripping, Program, 14(3) : 130-137
8. Zamir O., Etzioni O. Web document clustering: A feasibility demonstration. In Proceedings of the *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '98*, Melbourne, Australia, Aug. 24–28), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R., Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, 46–54.
9. http://snowball.tartarus.org/

Annotation:

Classification based on Suffix Tree phrases in comparison with Itemsets method

In this paper we present a text classification method using Suffix Tree (ST) phrases. We describe how to obtain ST-phrases from the training corpora, how to evaluate them and use them for text classification. Advantages and disadvantages of this approach are discussed and compared to the Itemsets method, which the Suffix Tree classification is based on. We also explain the way a threshold for multiclass classification is determined. We devote some time to examine the document length influence on classification effectiveness and also compare the impact of higher order Itemsets and ST-phrases in both methods. Of course, some comparison of the results obtained with other favourite text classification methods is provided at last.