Rozšíření bag-of-words modelu dokumentu: srovnání bigramů a 2-itemsetů

Roman Tesař[1], Massimo Poesio[2], Václav Strnad[1], Karel Ježek[1]

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

[2]Katedra Výpočetní techniky, Univerzita v Essexu, Wivenhoe Park, Colchester CO4 3SQ, Velká Británie
poesio@essex.ac.uk

Abstrakt. Jedním ze základních přístupů při kategorizaci textu je reprezentovat dokumenty jednotlivými slovy. Tento přístup je označován jako bag-of-words nebo také single words-based. Nicméně dalším obohacením této reprezentace je možné dosáhnout zlepšení výsledků klasifikace. V této práci jsme zaměřili svou pozornost na porovnání přínosu bigramů a 2-itemsetů, o které je rozšířen klasický bag-of-words model dokumentu. K experimentům využíváme standardní anglické textové korpusy Reuters-21578 a 20 Newsgroups. Ke klasifikaci je použit multinomial Naive Bayes, protože pro tuto klasifikační metodu a výše zmíněné korpusy byla publikována řada odborných publikací, se kterými naše dosažené výsledky srovnáváme. K výběru charakteristických položek (feature selection) využíváme 5 různých přístupů. Naše experimenty indikují, že použitím bigramů a 2-itemsetů je možné statisticky významně zvýšit úspěšnost klasifikace. Dále je v případě 2-itemsetů velmi důležité zvolit vhodný způsob výběru charakteristických položek. Na druhou stranu, v případě bigramů je možné dosáhnout zlepšení úspěšnosti klasifikace i použitím velmi jednoduchého přístupu. Z našich experimentů usuzujeme, že není příliš efektivní rozšiřovat reprezentaci textového dokumentu o 2-itemsety, protože pomocí bigramů je možné dosáhnout lepších výsledků a jejich generování je oproti 2-itemsetům méně náročné.

Klíčová slova: zpracování textu, výběr položek, klasifikace, model dokumentu, bigram, 2-itemset, n-gram, itemset, srovnání.

1 Úvod

Automatická klasifikace textu je jednou z významných úloh při zpracování přirozeného jazyka a představuje zařazování nově příchozích dokumentů do předem definovaných tříd. V této oblasti bylo zvěřejněno mnoho publikací a mnoho algoritmů bylo navrženo a zkoumáno. V současné době je tento problém stále více aktuální, protože objem dat nejen na internetu neustále roste a najít požadované informace mnohdy není snadné.

Nejen vhodný klasifikátor textu, ale i odpovídající reprezentace dokumentu ovlivňuje výslednou úspěšnost klasifikace. Ta je často velmi dobrá, i když je dokument reprezentován jen samostatnými slovy (tento přístup je označován jako single words-based nebo bag-of-words, dále jen BOW). Nicméně rozšířením BOW reprezentace

lze celkovou úspěšnost klasifikace dále zlepšit. Jednou z možností je použít k tomuto účelu n-gramy [15, 20, 5] nebo itemsety [12, 13, 23].

N-gram (v případě slov označovaný také jako fráze) je obecně sekvence *n* prvků z dané posloupnosti, které následují po sobě a zachovávají tedy své původní pořadí. Při zpracování textu mohou být prvky reprezentovány například písmeny. Tento přístup byl úspěšně použit v [1] při kategorizaci dokumentů ve vícejazyčném korpusu. V této práci [1] byly použity a srovnány multinomiální Naive Bayes klasifikátor, k-Nearest Neighbours (kNN) klasifikační algoritmus a Rocchio algoritmus. Z výsledků je patrné, že multinomiální Naive Bayes klasifikátor dosáhl nejkratšího času zpracování a nejvyšší úspěšnosti klasifikace.

Dalším, častěji používaným, přístupem při kategorizaci textu je použít sekvence slov místo písmen. V [15] byly dokumenty reprezentovány nejen jednotlivými slovy, ale i slovními n-gramy do délky 5. Bylo pozorováno, že n-gramy do délky 3 v kombinaci s Naive Bayes klasifikátorem mohou zlepšit úspěšnost klasifikace, přičemž nejvyšší přínos byl pozorován při obohacení BOW modelu dokumentu o 2-gramy (označované také jako bigramy). N-gramy pro n>3 již neměly vliv na výsledek klasifikace. V [10] bylo pozorováno, že delší slovní sekvence mohou výslednou úspěšnost klasifikace dokonce ovlivnit i negativně.

V [15] a [20] byly n-gramy úspěšně použity k obohacení BOW modelu dokumentu, byly tedy použity současně se samostatnými slovy. Další možností je použít n-gramy místo jednotlivých slov, kterými jsou n-gramy tvořeny. V takovém případě tedy BOW reprezentace neobsahuje slova, která jsou již v některém n-gramu obsažena. Jak bylo ukázáno v [20] a zmíněno v [5], tento přístup vede ve většině případů ke zhoršení výsledků klasifikace, zatímco při použití předchozího je možné dosáhnout znatelného zlepšení.

Při generování n-gramů mohou být místo všech slov uvažována například jen slovesa, podstatná jména nebo přídavná jména, u kterých se očekává větší důležitost. Jak bylo zmíněno v [5], výzkumníci při použití tohoto přístupu také dosáhli určitých zlepšení. Nicméně v [4] je zmíněn i fakt, že dosáhnout statisticky významné zlepšení úspěšnosti klasifikace rozšířením BOW modelu dokumentu o n-gramy není snadné a je většinou dosaženo jen při použití postupů, které nejsou "state-of-the-art". Tento fakt indikuje, že BOW model dokumentu je dostatečný a není jednoduché ho vylepšit.

Pro generování n-gramů existuje mnoho algoritmů, některé z nich mohou být nalezeny např. v [15, 10, 21].

Při zpracování textu jsou ke zlepšení úspěšnosti klasifikace často používány také itemsety. Obvykle reprezentují množinu *n* slov, které se současně vyskytují v dokumentu. Také obohacením BOW reprezentace dokumentů o itemsety se již zabývalo mnoho výzkumníků. V [23] byla použita multi-variate Bernoulliho verze Naive Bayes klasifikátoru a BOW reprezentace dokumentu byla obohacena o často se vyskytující itemsety. Výsledky klasifikace ukázaly určité zlepšení a také větší vyváženost při použití itemsetů. K jejich generování zde byl použit známý Apriori algoritmus, jehož několik modifikací bylo popsáno a úspěšně použito například v [2] nebo [12]. Podobně jako v [23], také v [12] bylo pozorováno zlepšení úspěšnosti klasifikace při použití itemsetů a popsáno zajímavé chování – pro klasifikační třídy s velkým počtem dokumentů se přesnost klasifikace zvyšovala s počtem přidaných itemsetů, zatímco pro menší třídy byl efekt opačný. Jedním z prezentovaných vysvětlení byla možnost, že obohacení BOW reprezentace dokumentů o příliš mnoho a příliš dlouhé itemsety způ-

sobí nevyváženost v trénovacích datech a chybné odhady pravděpodobností pro třídy s malým počtem dokumentů.

V [23] i v [12] byl použit práh minimálního výskytu pro výběr častých itemsetů určených k obohacení BOW modelu dokumentu. Protože zde nebyl určen maximální počet slov, které mohl itemset obsahovat, jejich počet v rámci jednoho itemsetu mohl být teoreticky velmi velký. Jak bylo ovšem v [12] poznamenáno, delší itemsety jsou méně časté a tudíž pravděpodobnost jejich výběru je malá.

Ke generování itemsetů je k dispozici mnoho algoritmů. Tato oblast je poměrně intenzivně zkoumána, protože se jedná o časově i paměťově náročný problém. Popis a srovnání několika možných přístupů, včetně Apriori algoritmu, je prezentován v [26].

Obecně řečeno, n-gramy jsou podmnožinou itemsetů. Při bližším zkoumání je patrné, že toto tvrzení neplatí zcela. V souladu s definicí itemsetů a algoritmy pro jejich generování se itemset může skládat jen ze vzájemně různých slov. Je však zřejmé, že n-gramy stejná slova obsahovat mohou a algoritmy určené k jejich generování to umožňují [10, 15, 21]. Nicméně prakticky je poměrně vzácné získat n-gram obsahující stejná slova, zvláště v případě delších n-gramů. Obvykle jsou reprezentovány speciálními výrazy, například "moucha tse tse" nebo "náboj dum dum". Tudíž n-gramy mohou být někdy nikoliv úplnou, ale jen částečnou podmnožinou itemsetů.

V této práci zaměřujeme svou pozornost na srovnání přínosu slovních n-gramů a itemsetů s ohledem na úspěšnost klasifikace textu. K tomu používáme multinomial Naive Bayes klasifikátor (viz [11, 19]) a dva často používané korpusy, na kterých testujeme, zda mohou být dosaženy lepší výsledky klasifikace rozšířením BOW modelu dokumentu o n-gramy nebo itemsety a jaký počet n-gramů nebo itemsetů zlepší výsledky klasifikace statisticky významně. Na základě pozorování uvedených v [15, 10] a dalších jsme se rohodli porovnat pouze vliv 2-gramů a 2-itemsetů, protože ty především přispívají ke zlepšení úspěšnosti klasifikace.

2 Generování n-gramů a itemsetů

Jak již bylo zmíněno dříve, n-gramy jsou označovány také jako slovní sekvence nebo fráze. Při zpracování přirozeného jazyka jsou rozlišovány statistické a syntaktické fráze [14, 7]. Autoři publikace [14] nepozorovali velké rozdíly v úspěšnosti klasifikace mezi syntaktickými a statistickými frázemi. Kromě toho, v [9] nebyl patrný žádný přínos při použití syntaktických frází na kolekci Reuters-21578. Protože i v naší práci jsme chtěli využít tento korpus, rohodli jsme se zaměřit jen na statistické fráze.

Při našich experimentech jsme ke generování n-gramů použili algoritmus prezentovaný v [21], který je modifikací Suffix Tree Clustering algoritmu popsaného v [24].

Pro generování itemsetů jsme využili upravený Apriori algoritmus (viz [26]).

3 Výběr charakteristických položek

Výběr charakteristických položek (feature selection) je technika sloužící k výběru podmnožiny takových položek dostupných z datové kolekce, které jsou pro klasifikaci nejpřínosnější. Tím je možné dimenzi kolekce podstatně zmenšit a odstranit položky

negativně ovlivňující klasifikaci. V této oblasti bylo mnoho publikováno, nejvhodnější pro naši úlohu, a také často citované, jsou [22] a [16].

Protože počet vygenerovaných 2-gramů a 2-itemsetů může být obrovský, zvolili jsme několik různých metod k výběru jen těch nejvhodnějších. Uvažujeme-li, že *f* reprezentuje položku z datové kolekce (v našem případě bigram nebo 2-itemset) a *c* klasifikační třídu, je v následujících kapitolách použita tato notace:

- *A* je počet kolikrát se *f* vyskytuje v dokumentech patřících do *c*
- *B* je počet kolikrát se *f* vyskytuje v dokumentech nepatřících do *c*
- *C* je počet kolikrát se *f* nevyskytuje v dokumentech patřících do *c*
- *D* je počet kolikrát se *f* nevyskytuje v dokumentech nepatřících do *c*
- *N* je počet všech dokumentů v kolekci
- *k* je počet všech klasifikačních tříd v kolekci

Některé metody pro výběr charakteristických položek berou v potaz podmíněné pravděpodobnosti jednotlivých slov, ze kterých se 2-itemset skládá [13]. V této práci jsme použili stejný přístup jako například v [23], kde jsou 2-itemsety považovány za nedělitelnou položku – stejně jako bigramy nebo samostatná slova.

3.1 Metoda χ^2 (CHI)

Metoda χ^2 určuje závislost mezi položkou *f* a třídou *c* [22] a obecně je známa její nespolehlivost pro nepříliš časté položky. Je definována jako

$$\chi^2(f,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{1}$$

Pokud je *f* nezávislá na *c*, výsledek vzorce je 0. Protože tento vzorec produkuje jednu hodnotu pro každou klasifikační třídu zvlášť, použili jsme pro stanovení celkového významu položky *f* v kolekci vždy jen nevyšší hodnotu

$$\chi^2_{\max}(f) = \max_{i=1}^{k}\left(\chi^2(f,c_i)\right) \tag{2}$$

Tento přístup se osvědčil nejen v [22], ale i později v [17]. Proto jsme ho aplikovali i na dále popsané metody produkující hodnoty pro každou klasifikační třídu zvlášť.

3.2 Mutual Information (MI)

Tato metoda stanovuje, kolik informace přináší přítomnost položky *f* o třídě *c*. Převzali jsme stejný způsob výpočtu jako byl použit v [22]:

$$MI(f,c) = \frac{P(f \wedge c)}{P(f) \times P(c)} \approx \frac{A \times N}{(A+C) \times (A+B)} \tag{3}$$

Pokud jsou *f* a *c* nezávislé, hodnota vzorce je 0. Metoda MI upřednostňuje nepříliš časté položky a může proto dosahovat horších výsledků [22].

3.3 Odds Ratio (OR)

OR je metoda podávající dobré výsledky zejména v kombinaci s Naive Bayes klasifikátorem [15, 16]. Protože vyžaduje informaci o výskytu *f* v dokumentech nepatřících do *c*, uvažujeme zde vždy všechny dokumenty kolekce ze všech tříd kromě *c*. Pokud

$$P(f \mid c) = \frac{A}{A+C} \quad , \quad P(f,\bar{c}) = \frac{B}{B+D} \ ,$$

potom je OR definováno jako

$$OR(f,c) = \frac{P(f,c)\cdot\left(1-P(f,\bar{c})\right)}{\left(1-P(f,c)\right)\cdot P(f,\bar{c})}. \tag{4}$$

Ze vzorce (4) je patrné, že pokud $P(f,c)=1$ nebo $P(f,\bar{c})=0$, tedy v případech, kdy se *f* vyskytuje pouze v dokumentech třídy *c*, potom $OR(f,c)=\infty$. V takových případech uvažujeme dodatečné ohodnocení pomocí $DF(f \mid c)$ (viz sekce 3.5).

3.4 Information Gain (IG)

IG určuje obecnou významnost položky *f* s ohledem na všechny klasifikační třídy a je stanoven jako

$$IG(f) = -\sum_{i=1}^{k} P(c_i)\log P(c_i) + P(f)\sum_{i=1}^{k} P(c_i \mid f)\log P(c_i \mid f) + P(\bar{f})\sum_{i=1}^{k} P(c_i \mid \bar{f})\log P(c_i \mid \bar{f}), \tag{5}$$

kde $P(c_i)$, $P(c_i \mid f)$ a $P(c_i \mid \bar{f})$ jsou určeny následujícím způsobem

$$P(c_i) = \frac{A+C}{N}, \quad P(c_i \mid f) = \frac{A}{A+B}, \quad P(c_i \mid \bar{f}) = \frac{C}{C+D}.$$

Tuto definici jsme převzali z [22], kde IG podával dobré výsledky ve srovnání s jinými metodami. Nicméně, jak bylo uvedeno v [17], položky s nižší entropií mají nižší IG ohodnocení, ačkoliv mohou být silně korelované s *c*. To je způsobeno faktem, že IG stoupá také s nárůstem entropie položky *f*.

3.5 Document Frequency (DF)

Prahování pomocí DF je jedním z nejjednodušších přístupů pro výběr charakteristických položek a je obvykle považováno za "ad hoc" přístup. DF určuje počet dokumentů třídy *c*, ve kterých se vyskytuje položka *f* a může být stanoven jako

$$DF(f \mid c) = A. \tag{6}$$

Navzdory své jednoduchosti bylo v [18] a [22] ukázáno, že DF je důležitou metrikou, protože upřednostňuje časté položky, což se zdá být významnou charakteristikou.

4 Experimenty

Jak již bylo zmíněno, pro porovnání vlivu bigramů a 2-itemsetů byla zvolena multinomiální verze Naive Bayes klasifikátoru. Pro obě použité datové kolekce (viz sekce 4.1) jsme vždy uvažovali výsledek klasifikace s použitím všech slov jako základ, oproti kterému jsme porovnávali přínos určitého počtu bigramů a 2-itemsetů, které byly vybrány přístupy popsanými v kapitole 3. Protože některé z těchto přístupů (např. MI a OR) zvýhodňují málo se vyskytující položky, které nemohou výrazněji ovlivnit klasifikaci, uvažovali jsme u obou kolekcí jen bigramy a 2-itemsety s určitým minimálním počtem výskytů. K ověření statistické významnosti výsledků jsme použili neparametrický McNemarův test (viz [8]) a 99% interval spolehlivosti.

4.1 Datové kolekce

4.1.1 Reuters-21578

Bylo prokázáno (viz [5]), že tato kolekce je příkladem "simple" kolekce. Tím se rozumí, že s použitím několika málo vhodně vybraných slov je možné se velmi přiblížit nejlepšímu výsledku klasifikace, který byl prozatím na této kolekci dosažen. Nicméně z [19] je patrné, že použití všech slov ke klasifikaci stále poskytuje nejlepší výsledek.

Pro naše experimenty jsme zvolili “ModApté” verzi kolekce Reuters-21578, ze které jsme využili jen 10 nejčastěji zastoupených tříd, kdy je celkový počet trénovacích dokumentů 6490, testovacích 2545 a 1.1 průměrný počet tříd na dokument. Stejný přístup byl aplikován např. i v [11, 2, 19]. Z kolekce jsme v rámci předzpracování odstranili stopslova a čísla. Všechna písmena byla následně převedena na malá. Stemming ani lematizace nebyly použity. Po fázi předzpracování obsahovala trénovací sada dokumentů 22208 různých slov s průměrným počtem 73 slov na dokument a průměrným počtem 47 různých slov na dokument. Protože tato kolekce obsahuje dokumenty zařazené do více tříd současně, použili jsme Naive Bayes klasifikátor pro každou ze tříd zvlášť, což je obvyklý přístup, označovaný také jako “one-vs-rest”.

4.1.2 20 Newsgroups

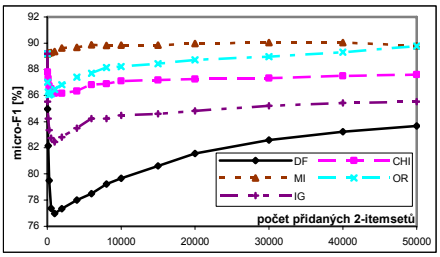
Oproti předchozí není tato kolekce "simple" kolekcí (viz [5]) - tedy každé slovo, které je použito ke klasifikaci, přispívá ke zlepšení výsledku klasifikace. Pro naše účely jsme zvolili "bydate" verzi 20 Newsgroups kolekce, která obsahuje 11314 trénovacích a 7532 testovacích dokumentů. I zde jsme aplikovali stejný postup předzpracování jako na předchozí kolekci, čímž jsme získali 100354 unikátních slov v trénovací množině dokumentů, průměrný počet slov na dokument 153 a průměrně 99 různých slov v dokumentu.

4.2 Výsledky testů na kolekci Reuters-21578

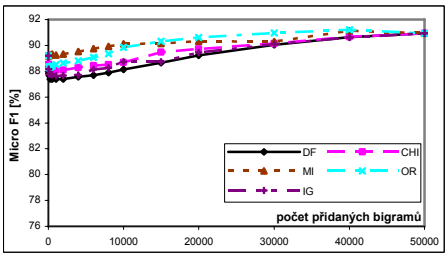
U této kolekce jsme pro generování bigramů a 2-itemsetů uvažovali jen slova s četností 2 a více, což zredukovalo jejich počet z 22 208 na 12 127. Na vygenerované bigramy a 2-itemsety jsme aplikovali stejný práh, což zredukovalo počet itemsetů

z 4 778 151 na 1 870 619 a počet bigramů z 207 018 na 49 872. Z těchto hodnot je patrné, že počet řídce se vyskytujících položek v kolekci Reuters-21578 je značný. To může být jedním z důvodů, proč i jen několik málo nejčastějších slov poskytuje dobré výsledky při klasifikaci a zahrnutí dalších už nepřináší příliš výrazné zlepšení.

Hodnota 0 na ose X (pro Obr. 1 až 4) představuje situaci, kdy byla ke klasifikaci použita pouze slova. Dosaženo zde bylo hodnoty *micro-F1*= 89.20% (viz spodní část Tab. 1 a 2), kterou chápeme jako základ a se kterou dále srovnáváme. Jak je patrné, kromě MI vykazují ostatní přístupy ke generování charakteristických položek významný pokles výsledků klasifikace, zvláště při použití malého počtu přidaných 2-itemsetů (Obr. 1).



Obr. 1. Závislost micro-F1 na počtu přidaných 2-itemsetů pro kolekci Reuters-21578



Obr. 2. Závislost micro-F1 na počtu přidaných bigramů pro kolekci Reuters-21578

Tabulka 1. Nejlepší výsledky pro 2-itemsety na kolekci Reuters-21578

	přidané 2-itemsety	*micro F1*	*macro F1*	*BEP*	odebrané 2-itemsety	*micro F1*	*macro F1*	*BEP*
DF	0	89.20	80.21	89.84	---	---	---	---
CHI	0	89.20	80.21	89.84	---	---	---	---
MI	40000	90.06	82.80	90.31	---	---	---	---
OR	50000	89.79	81.75	90.35	-25000	90.44	82.33	90.79
IG	0	89.20	80.21	89.84	---	---	---	---
	0 (základ)	*89.20*	*80.21*	*89.84*				

Tabulka 2. Nejlepší výsledky pro bigramy na kolekci Reuters-21578

	přidané bigramy	*micro F1*	*macro F1*	*BEP*	odebrané bigramy	*micro F1*	*macro F1*	*BEP*
DF	49872	90.93	81.65	91.24	-8000	91.39	82.84	91.47
CHI	49872	90.93	81.65	91.24	-2000	91.35	82.11	91.5
MI	40000	91.13	82.64	91.36	---	---	---	---
OR	40000	91.21	81.92	91.48	-2000	91.47	82.27	91.67
IG	49872	90.93	81.65	91.24	-5000	91.41	82.17	91.5
	0 (základ)	*89.20*	*80.21*	*89.84*				

Při bližším zkoumání se potvrdilo, že MI upřednostňuje málo časté položky, které obecně nemohou významně ovlivnit výsledek klasifikace. Nicméně na základě vzorce 3 byly vybrány 2-itemsety z tříd s menším počtem dokumentů, což pomohlo v jejich lepší identifikaci. Oproti tomu DF, CHI a IG upřednostňují spíše časté položky, což se nezdá být vhodný přístup u této kolekce. Příčinou může být fakt, že se jedná o "simple" kolekci (viz sekce 4.1), kde jednotlivá slova postačují k rozlišení jednotlivých tříd a rozšířením tohoto single words-based modelu o příliš časté 2-itemsety, které mají výrazný vliv na Naive Bayes klasifikátor, může být model dokumentů v kolekci zkreslen a tím i zhoršen výsledek klasifikace. Jak již bylo zmíněno v [22], DF, CHI a IG jsou silně korelovány a jejich chování je tedy podobné. OR má zajímavé chování. Pro malý počet přidaných 2-itemsetů se výsledek klasifikace zhoršuje. Nicméně, na rozdíl od CHI a IG, použitím více 2-itemsetů s nižším ohodnocením znatelně zlepšuje výsledek klasifikace. Důvodem může být fakt, že 1000 nejlépe ohodnocených 2-itemsetů pomocí OR je převážně jen ze třídy *earn*, která obsahuje nejvíce dokumentů. Dále však již třída *earn* nedominuje a přítomny jsou především 2-itemsety z tříd s méně dokumenty, což je v kontrastu především vzhledem k DF a IG. Navíc, jak bylo ukázáno v [16], OR upřednostňuje položky s častějším výskytem v dané třídě před

položkami, které se v dané třídě nevyskytují, což se zdá být vhodný přístup při použití Naive Bayes klasifikátoru.

Jak je patrné z Obr. 1, pouze 2-itemsety vybrané pomocí OR a MI byly schopny překročit základní *micro-F1* hodnotu 89.20%. Tato zlepšení ale nejsou statisticky významná. Levá strana Tab. 1 obsahuje nejlepší hodnoty z charakteristik na Obr. 1.

V případě bigramů je situace odlišná. Jak je patrné z Obr. 2, všechny přístupy pro výběr charakteristických položek překonaly základní *micro-F1* hodnotu 89.20%. I zde došlo nejprve v případě DF, CHI, OR a IG k poklesu úspěšnosti klasifikace, avšak méně výraznému. Navzdory faktu, že bigramy mají obecně menší počet výskytů v kolekci než 2-itemsety a je jich podstatně méně, jejich přínos při klasifikaci je větší. V levé části Tab. 2 jsou nejlepší výsledky z charakteristik na Obr. 2. Všechna zlepšení oproti základu 89.20% jsou statisticky významná.

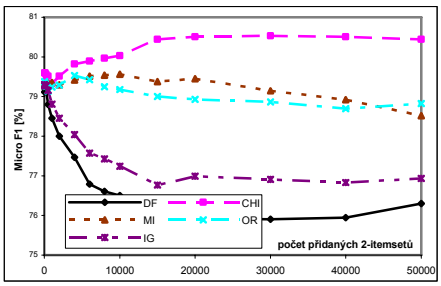
S ohledem na fakt, že příliš časté 2-itemsety a bigramy u kolekce Reuters-21578 znatelně zhoršují výsledky klasifikace, jsme se rozhodli určitý počet těch nejlépe ohodnocených jednotlivými přístupy neuvažovat. Pro každý z přístupů jsme ponechali počet 2-itemsetů a bigramů, u kterého bylo dosaženo nejlepších výsledků (viz Tab. 1 a 2 vlevo), a odebírali postupně nejlépe ohodnocené položky. Výsledky pro 2-itemsety a bigramy jsou uvedeny v Tab. 1 a 2 vpravo.

Z Tab. 1 je patrné, že pouze u OR došlo ke zlepšení při odebrání 25 000 nejlépe ohodnocených 2-itemsetů (o ostatních metod žádné zlepšení oproti základní *micro-F1* hodnotě nenastalo), což jsme vzhledem k charakteristice na Obr. 1 očekávali. Zlepšení z 89.79% na 90.44% je statisticky významné.

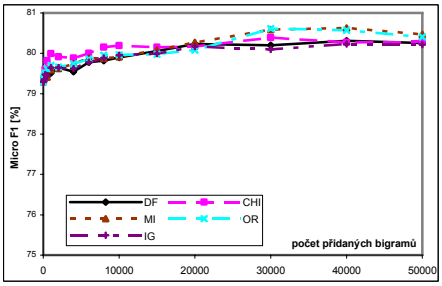
Odebráním menšího počtu nejlépe ohodnocených bigramů došlo (viz Tab. 2), kromě MI, u všech ostatních přístupů ke zlepšení výsledků klasifikace (nicméně statisticky nevýznamnému). Ačkoliv jsou rozdíly mezi DF, CHI, OR a IG malé, OR i zde podává nejlepší výsledky. Podobný závěr byl učiněn i v [16], kde Naive Bayes klasifikátor vykazoval v kombinaci s OR nejlepší výsledky. Na základě nám známých publikovaných prací je doposud nejlepším výsledkem dosaženým na kolekci Reuters-21578 hodnota break-even-point (BEP = bod, kde jsou hodnoty přesnosti a úplnosti klasifikace stejné) 92% a byla dosažena pomocí SVM klasifikátoru a MI přístupu (viz [9]). Náš výsledek *BEP*=91.67% dosažený pomocí Naive Bayes klasifikátoru a OR přístupu rozšířením single words-based modelu dokumentu o vhodný počet bigramů je poměrně blízko této hodnotě. Je to bezpochyby způsobeno faktem, že Reuters-21578 je příkladem "simple" kolekce (viz sekce 4.1.1).

4.3 Výsledky testů na kolekci 20 Newsgroups

Tato kolekce obsahovala po předzpracování 100 345 unikátních slov. Při výběru charakteristických položek jsme uvažovali jen slova s minimálním počtem výskytů 5 a více, což zredukovalo jejich počet na 22 499. Z nich jsme získali 734 673 bigramů. Počet 2-itemsetů není díky implicitním optimalizačním procesům k dispozici. Protože počet položek je poměrně velký, aplikovali jsme, podobně jako u předchozí kolekce, na vygenerované bigramy a 2-itemsety minimální práh výskytu – uvažovali jsme jen položky s výskytem 4 a více. Tento krok zredukoval počet bigramů na 50 7563 a počet 2-itemsetů na 5 521 698. Obr. 3 prezentuje *micro-F1* charakteristiky jednotlivých přístupů pro výběr 2-itemsetů. Základní *micro-F1* hodnota, ze které jsme vycházeli a kterou jsme získali použitím všech unikátních slov ke klasifikaci, je 79.3%.



Obr. 3. Závislost micro-F1 na počtu přidaných 2-itemsetů pro kolekci 20 Newsgroups



Obr. 4. Závislost micro-F1 na počtu přidaných bigramů pro kolekci 20 Newsgroups

Tabulka 3. Nejlepší výsledky pro 2-itemsety na kolekci 20 Newsgroups

	přidaných 2-itemsetů	*micro F1*	*macro F1*
DF	0	79.30	76.82
CHI	30000	80.54	79.02
MI	10000	79.57	77.20
OR	4000	79.53	77.33
IG	0	79.30	76.82
	0 (baseline)	*79.30*	*76.82*

Tabulka 4. Nejlepší výsledky pro bigramy na kolekci 20 Newsgroups

	přidaných bigramů	*micro F1*	*macro F1*
DF	40000	80.31	78.35
CHI	30000	80.39	78.31
MI	40000	80.63	78.53
OR	30000	80.62	78.40
IG	40000	80.23	78.23
	0 (baseline)	*79.30*	*76.82*

U této kolekce se zdá být CHI nejvhodnějším přístupem pro generování 2-itemsetů (viz Obr. 3). Přibližně prvních 1000 2-itemsetů způsobuje pokles úspěšnosti klasifikace, který je ovšem méně výrazný než u předchozí kolekce a dále následuje postupné zlepšování výsledků. MI a OR se chovají podobně a především z důvodu výběru převážně méně častých položek neovlivňují výrazně průběh klasifikace. DF a IG nepodávají dobré výsledky z důvodu výběru především 2-itemsetů, které jsou velmi časté v několika třídách současně. IG navíc nerozlišuje mezi přítomností a absencí položky ve třídě, což patrně není vhodný přístup pro kombinaci s multinomiální verzí Naive Bayes klasifikátoru, který bere v potaz jen výskyt položky. Podobný závěr byl předložen také v [16]. U této kolekce je rozdíl mezi nejhorší a nejlepší dosaženou hodnotou při využití 2-itemsetů znatelně menší než u kolekce předchozí. Důvodem může být fakt, že 20 Newgroups má rovnoměrnější distribuci dokumentů v jednotlivých třídách a není příkladem "simple" kolekce. Nejlepší dosažené hodnoty z Obr. 3 jsou k dispozici v Tab. 3[2]. Pouze v případě CHI se jedná o statisticky významné zlepšení.

Na Obr. 4 jsou k dispozici charakteristiky pro bigramy. Jak je patrné, nedochází zde při využití nejlépe ohodnocených bigramů k poklesu úspěšnosti klasifikace. CHI se zdá být vhodné při využití jen menšího počtu nejlépe ohodnocených bigramů, zatímco OR podává dobré výsledky především při využití jejich většího počtu. CHI zde pravděpodobně vhodně kombinuje výběr položek nejlépe charakterizujících jednotlivé třídy s četnostmi jejich výskytů, což přináší zlepšení úspěšnosti klasifikace i při využití jen malého počtu bigramů.

MI a OR stále vybírají většinou méně časté bigramy, což v tomto případě přináší také zlepšení, zejména při použití 30 000 a 40 000 bigramů.

[2] V Tab. 3 a 4 nejsou uvedeny hodnoty *BEP*, protože v kolekci 20 Newsgroups patří každý dokument jen do jediné třídy a tedy *micro-F1* a *BEP* hodnoty jsou stejné.

DF je svým jednoduchým přístupem také použitelné v této klasifikační úloze, což indikuje, že bigramy pravděpodobně způsobují při reprezentaci dokumentů menší zkreslení než 2-itemsety. Těch je obvykle vygenerováno mnohem více a je tedy náročnější vybrat jen ty vhodné. Korelace mezi IG a DF je i u této kolekce patrná.

Nejlepší dosažené výsledky pro charakteristiky z Obr. 4 jsou uvedeny v Tab. 4. Všechna zde uvedená zlepšení jsou oproti základu 79.30% statisticky významná.

Podobný základ, jenž jsme uvažovali u této kolekce, byl uvažován i v [19] a naše nejlepší dosažená hodnota *micro-F1*=80.63% (viz Tab. 4) je srovnatelná s [3]. Nicméně v [6] byl také použit Naive Bayes klasifikátor s několika odlišnými modely dokumentu a bylo zde dosaženo úspěšnosti klasifikace *micro-F1*=82.21%. Při použití klasifikátoru SVM s jádrem NGD byla v [25] publikována úspěšnost klasifikace 84.61% na kolekci 20 Newsgroups ("bydate" verze).

5 Závěr

V naší práci jsme se zaměřili na srovnání přínosu 2-itemsetů a bigramů použitých k obohacení modelu dokumentu založeném jen na slovní reprezentaci (BOW) při klasifikaci textu. Výsledky na dvou anglických kolekcích indikují, že 2-itemsety jsou příliš obecné a vyvolávají často při klasifikaci nežádoucí zkreslení. Velký počet vygenerovaných 2-itemsetů společně s nevyváženým zastoupením klasifikačních tříd v datové kolekci vyžaduje pečlivý výběr metody pro výběr charakteristických položek. Naproti tomu, bigramy se zdají být vhodnější pro úlohu klasifikace. Výběr charakteristických bigramů není tolik závislý na použitém přístupu. Uvážíme-li, že Apriori algoritmus a jiné určené ke generování itemsetů jsou většinou časově i paměťově mnohem náročnější než algoritmy ke generování bigramů, nezdá se být příliš efektivní rozšiřovat při klasifikaci model dokumentu o 2-itemsety. Stejného, případně lepšího, výsledku lze dosáhnout použitím bigramů.

V budoucnu se chystáme stejné experimenty zopakovat na kolekcích v českém jazyce, kde očekáváme podobné závěry.

Tato práce byla částečně podporována z prostředků Národního Programu Výzkumu II, projekt 2C06009 (COT-SEWing).

Reference

1. J.J.G. Adeva, R.A. Calvo and D.L. de Ipiña. Multilingual Approaches to Text Categorisation. UPGRADE: *The European Journal for the Informatics Professional*, Vol. VI, No. 3, pp. 43 - 51, June 2005.
2. M.L. Antonie and O.R. Zaiane. Text document categorization by term association. *In Proc. of the IEEE 2002 International Conference on Data Mining*, pages 19–26, Maebashi City, Japan, 2002.
3. L. Baoli, L. Qin and Y. Shiwen. An adaptive k-nearest neighbor text categorization strategy. *ACM Transactions on Asian Language Information Processing (TALIP)*, volume 3, pp: 215 - 226, 2004.

4. R. Bekkerman, R. El-Yaniv, N. Tishby and Y. Winter. Distributional Word Clusters vs. Words for Text Categorization. *Journal of Machine Learning Research*, 3:1183-1208, 2003.
5. R. Bekkerman and J. Allan. Using Bigrams in Text Categorization. CIIR Technical Report IR-408, 2004.
6. A. Bratko and B. Filipič. Exploiting Structural Information in Semistructured Document Classification. *Proc. 13th International Electrotechnical and Computer Science Conference, ERK 2004*, 2004.
7. M. F. Caropreso, S. Matwin and F. Sebastiani. Statistical phrases in automated text categorization. Technical Report IEI-B4-07-2000, Istituto di Elaborazione dell'Informazione, Pisa, Italy, 2000.
8. T.G. Dietterich. Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, Vol. 10, pp. 1895-1923, 1998.
9. S.T. Dumais, J. Platt, D. Heckerman and M. Sahami. Inductive learning algorithms and representations for text categorization. *In Proceedings of ACM-CIKM98*, pp. 148-155, 1998.
10. J. Fürnkranz. A study using n-gram features for text categorization. Technical Report OEFAI-TR-9830, Austrian Institute for Artificial Intelligence, Vienna, Austria, 1998.
11. A. McCallum and K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. *In AAAI/ICML-98 Workshop on Learning for Text Categorization*, Technical Report WS-98-05, AAAI Press, pp. 41-48, 1998.
12. D. Meretakis and B. Wüthrich. Extending Naive Bayes classifiers using long itemsets. *In Proc. 5th ACM-SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, San Diego, USA, pp. 165-174, 1999.
13. D. Meretakis , D. Fragoudis , H. Lu and S. Likothanassis. Scalable association-based text classification. *Proceedings of the ninth international conference on Information and knowledge management*, pp. 5-11, McLean, Virginia, United States, November 2000.
14. M. Mitra, C. Buckley, A. Singhal and C. Cardie. An analysis of statistical and syntactic phrases. *In Proceedings of RIAO-97, 5th International Conference "Recherche d'Information Assistee par Ordinateur"*, pp. 200-214, Montreal, CA, 1997.
15. D. Mladenic and M. Grobelnik. Word sequences as features in text-learning. *In Proc. 17th Electrotechnical and Computer Science Conference (ERK98)*, Slovenia, 1998.
16. D. Mladenic and M. Grobelnik. Feature Selection for Unbalanced Class Distribution and Naive Bayes. *In Proceedings of the 16th International Conference on Machine Learning*, Morgan Kaufmann, pp. 258-267, 1999.
17. V. Pekar, M. Krkoska and S. Staab. Feature Weighting for Co-occurrence-based Classification of Words. *In Proceedings of the 20th Conference on Computational Linguistics, COLING-2004*, August 2004.
18. M. Rogati and Y. Yang. High-performing feature selection for text classification. *Proceedings of the eleventh international conference on Information and knowledge management*, McLean, Virginia, USA, November 2002.

19. K.M. Schneider. A New Feature Selection Score for Multinomial Naive Bayes Text Classification Based on KL-Divergence. *42nd Meeting of the Association for Computational Linguistics (ACL 2004)*, pp. 186-189, 2004.
20. Ch.M. Tan, Y.F. Wang and Ch.D. Lee. The use of bigrams to enhance text categorization. *Information Processing and Management: an International Journal*, v.38 n.4, p.529-546, July 2002.
21. R. Tesar, D. Fiala, F. Rousselot and K. Jezek. A comparison of two algorithms for discovering repeated word sequences. *WIT Transactions on Information and Communication Technologies*, Vol. 35, pp. 121 - 131, 2005.
22. Y. Yang and J.O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, pp.412-420, July, 1997.
23. Z. Yang, Z. Lijun, Y. Jianfeng and L. Zhanhuai. Using association features to enhance the performance of Naive Bayes text classifier. *Fifth International Conference on Computational Intelligence and Multimedia Applications*, *ICCIMA 2003*, pp. 336-341, 2003.
24. O. Zamir and O. Etzioni. 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), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R., Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, pp. 46–54, 1998.
25. D. Zhang, X. Chen and W.S. Lee. Text classification with kernels on the multinomial manifold. *Proceedings of the 28th international ACM SIGIR conference on Research and development in information retrieval*, Brazil, pp. 266-273, 2005.
26. Z. Zheng, R. Kohavi and L. Mason. Real world performance of association rule algorithms. *In Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, August 2001.

Annotation:

Extending the Bag-of-Words Document Model: A Comparison of Bigrams and 2-Itemsets

In this paper, we compare the performance improvement in terms of classification accuracy when bigrams and 2-itemsets are used to extend the single words-based document representation on two standard text corpora: Reuters-21578 and 20 Newsgroups. The conclusion is that it is not very effective to extend document representation with 2-itemsets. Bigrams achieve better results and discovering them is less resource-consuming.

This paper is a shortened, Czech version of the following publication
Tesar R., Poesio M., Strnad V., Jezek K.: "Extending the Single Words-Based Document Model: A Comparison of Bigrams and 2-Itemsets". *In Proceedings of the 2006 ACM Symposium on Document Engineering (DocEng'06)*, Amsterdam, Netherlands, ACM press, ISBN 1-59593-515-0, pages 138-146, September 2006. (http://doi.acm.org/10.1145/1166160.1166197)