Sumarizace textů

Karel JEŽEK[1], Josef STEINBERGER[2]

[1]*Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni*
Univerzitní 8, 306 14 Plzeň
Karel Jezek <jezek_ka@kiv.zcu.cz>
[2]*European Commission Joint Research Centre, IPSC Ispra*
T.P. 267, 21027 Ispra (VA), Italy
Josef Steinberger <josef.steinberger@jrc.ec.europa.eu>

Abstrakt. Jsme zahlcováni stále větším množstvím informací. Proto je v současné době velká pozornost věnována výzkumu a vývoji redukčních metod, které zachovávají co nejvyšší informační hodnotu redukovaných dat. Úlohy tohoto typu známe pod názvem vytváření abstraktů, extraktů či sumarizace a lze je aplikovat na data všech možných forem. V tomto výkladu se zaměříme na data textového tvaru, který lze stále ještě považovat za základní formu pro sdílení informací. Budou popsány principy a možnosti jak klasických sumarizačních metod, tak i metod založených na moderních algebraických postupech. Věnujeme také pozornost způsobu řešení těch úloh, které navazují na základní sumarizaci jednoho dokumentu. Jedná se zajména o úlohy sumarizace vícedokumentové a aktualizační, kde je nutné řešit navíc problém nežádoucí redundantní informace ve výsledku. Součástí výkladu bude i popis způsobů hodnocení kvality sumarizace a prezentování výsledků našeho vlastního výzkumu v této oblasti.

Klíčová slova: sumarizace, sumarizace textu, vícedokumentová sumarizace, aktualizační sumarizace, redukce dat, singulární dekompozice, latentní sémantická analýza.

1 Úvod

Snadné a levné zpřístupnění informací prostřednictvím zejména WWW, způsobilo jejich dříve nepředstavitelný nárůst a podnítilo zvýšený zájem o prostředky usnadňující jejich zvládnutí. Většina na webu přístupných stránek je tvořena textem. Proto je zvýšené úsilí věnováno metodám, které zhušťují textem poskytované informace, zkracují čas potřebný pro seznámení se s hlavními myšlenkami prezentovanými textem nebo alespoň rychle zprostředkují povědomí o tématech, kterým se text věnuje. Výsledné shrnutí musí pomocí výběru nebo s použitím zobecnění informovat o obsahu a závěrech originálního textu. Přesto, že s pojmem sumarizace všichni běžně pracujeme, považujeme za vhodné uvést v úvodu jeho definici. Vytváření souhrnu (sumarizaci) můžeme definovat např. jako:

- Vytvoření stručné a přesné reprezentace obsahu dokumentu,
- Vyjmutí nejdůležitější informace ze zdrojového textu, která jej zestručňuje pro účely a úlohy uživatele.

Sumarizace je úlohou, která je řešena od nepaměti. Až do vzniku počítačů samozřejmě manuálním způsobem, ale ani použití počítačů k jejímu řešení není horkou novinkou. Prvé pokusy se datují do poloviny minulého století. Za prvou publikaci pojednávající o

počítačové sumarizaci textu lze považovat Luhnovu práci [19], inspirovanou již tehdy informačním přetížením. Jím navržená metoda používá pro výběr vět do souhrnu frekvenci termů (slov nebo i frází). Jiným významným přínosem byla o něco později uvedená Edmunsonova práce [7], která vycházela z poznatku, že věty s nejvyšší informační hodnotou se obvykle vyskytují na začátku dokumentu. V devadesátých letech se v řešení sumarizačního problému začaly uplatňovat metody umělé inteligence. Uveďme např. [17], popisující učící se systém založený na Naive Bayes klasifikátoru, který je trénován na korpusu dvojic dokument – souhrn.

Současné výkonné počítače, nástup WWW a nové poznatky využitelné při řešení sumarizační úlohy, spolu s potřebou řešit dnešní informační přetížení jsou podnětem, na který reagovala řada výzkumných pracovišť pravidelně porovnávajících výsledky své práce na specializovaných konferencích jako je Document Understanding Conference (DUC), nově Text Analysis Conference (TAC). Začaly se řešit úlohy sumarizace více dokumentů, multimediálních dokumentů a sumarizace aktualizační, která poskytuje informace zohledňující předchozí znalosti uživatele a podává mu jen nové informace, vypouštěním informací obsažených v dokumentech, se kterými již byl dříve seznámen.

Důležitou související úlohou je vyhodnocování kvality sumarizace a jejích kvantitativních vlastností. Tyto vlastnosti můžeme měřit podle následujících, vzájemně ortogonálních hledisek:

- Sémantické informativnosti,
- Souvislosti textu,
- Kompresního poměru.

Sémantickou informativností rozumíme míru možnosti zrekonstruovat ze souhrnu původní text. Souvislostí rozumíme míru s jakou na sebe navazují jednotlivé části souhrnu a vytváří tak integrovaný výsledný text. Kompresní poměr je podílem délky souhrnu a délky originálu.

Způsob hodnocení může vycházet z porovnávání výsledného souhrnu s původním textem, s ručně vytvořeným souhrnem nebo se souhrnem vytvořeným jiným sumarizačním systémem. Hodnotící metody mohou být rozděleny na:

- Přímé metody, které jsou založené přímo na analýze souhrnu a jeho porovnání s originálem co do míry tematické obsažnosti, souvislosti, čitelnosti, gramatiky apod. Porovnávat výsledek je možné i s ručně vytvořeným abstraktem (od autora originálu nebo od profesionálního abstraktora),
- Nepřímé metody, které jsou založené na míře použitelnosti souhrnu pro zadaný účel. Tím může být např. klasifikační úloha, filtrování, vyhledávání nebo odpovídání na dotazy. Kvalita souhrnu pak může být určena kvantitativními ukazateli jako jsou třeba přesnost a úplnost výběru podle souhrnu v porovnání s výběrem podle originálu nebo „ideálního“, ručně konstruovaného souhrnu.

Poznamenejme, že pojem „ideálního“ souhrnu je pouhou fikcí a pracovat s ním je třeba obezřetně.

Další části příspěvku jsou uspořádány takto: Následující část zavádí terminologii a obecně pojednává o jednotlivých způsobech automatické sumarizace. V další části se seznámíme s tradičními metodami, které jsou založeny na heuristických či statistických postupech a vznikly v minulém miléniu, dále se věnujeme novějším sumarizačním postupům. V páté kapitole se budeme věnovat algebraickým způsobům sumarizace, které jsou pozoruhodné používáním metod maticové faktorizace. Šestá kapitola je věnována úlohám, které navazují na jednodokumentovou sumarizaci a jsou předmětem zájmu současného výzkumu. Uvedeme rovněž výsledky některých vlastních prací. Sedmá část seznamuje se způsoby hodnocení kvality sumarizace. Následuje závěr, s výhledem na další možný výzkum.

2 Typy sumarizátorů

Existuje několik navzájem nezávislých hledisek, která mohou být použita k zavedení taxonomie sumarizátorů. Uveďme ta nejčastěji používaná hlediska a z nich vycházející členění.

- Forma souhrnu:
  - Extrakt je souhrn zcela tvořený sekvencemi slov, které jsou okopírovány z původního dokumentu. Jako kopírované úseky mohou být použity fráze, věty nebo i celé odstavce originálu. Jak lze předpokládat, extrakty trpí chabou souvislostí zařazených úseků, způsobenou zejména častým opomíjením anaforických vztahů. Výběr vět může být proveden bez ohledu na kontext, výsledek bývá nevyvážený a nesourodý.
  - Abstrakt je souhrn, který nemusí obsahovat a většinou neobsahuje sekvence slov z originálního textu. V současné době se stále ještě jedná o úlohu, která je pro počítačové zpracování obtížně řešitelná. Vyžaduje analýzu vstupního textu včetně sémantické analýzy a následnou syntézu, generující věty v přirozeném jazyce.
- Úroveň zpracování souhrnu:
  - Povrchní přístupy, ve kterých jsou informace reprezentovány prostřednictvím povrchních vlastností a jejich kombinacemi. Povrchními vlastnostmi jsou např. pozičně významné termy (vžilo se používat slovo *term* místo češtějšího *termín*), frekvenčně významné termy, termy specifické pro zpracovávanou doménu nebo termy obsažené v uživatelově dotazu. Jejich výsledkem je extrakt.
  - Hlubší přístupy mohou produkovat extrakt nebo i abstrakt. K určení významných částí textu využívají jeho sémantické zpracování, zjišťují textové jednotky a jejich vzájemné vztahy jako jsou tezaurové relace, syntaktické relace apod. Mohou využívat informací o stavbě textu a rétorické struktuře, případně i hypertextových značek.
- Účel, pro který je souhrn vytvářen:
  - Hodnotící souhrny, do kterých lze začlenit kritiky, recenze, posudky. Jejich charakteristickým rysem je, že vyjadřují mínění autora souhrnu o daném dokumentu. Tato okolnost zatím prakticky vylučuje hodnotící souhrny ze skupiny automaticky generovatelných.
  - Indikativní souhrny dávají zkrácenou formou informaci o hlavních tématech dokumentu, zachovávají jeho nejpodstatnější části. Měly by umožnit uživateli rozhodnout, zda čtení celého textu bude pro něj dostatečně přínosné. Jsou proto často využívány ve výstupech vyhledávacích systémů, kde nahrazují originální texty dokumentů. Jejich obvyklá délka bývá do 10% úplného textu.
  - Informativní souhrny nahrazují originální dokument poskytnutím jeho stručného obsahu. Při zkrácení původního textu o 70-80%, může si souhrn zachovat i důležité detaily originálu. Míra informování čitatele by měla postačovat pro zběžné seznámení s tématem a vyhnout se tak čtení celého dokumentu.
- Podle uživatelů můžeme souhrny rozdělit např. na:
  - Obecné souhrny, které jsou určeny pro širokou třídu čtenářů, s různými zájmovými oblastmi. Pro obecný souhrn jsou důležitá všechna v dokumentu obsažená témata.

  - Souhrny založené na dotazu, jejich obsah je vytvořen tak, aby vybral z dokumentu informace relevantní k dotazu uživatele.
  - Tematicky zaměřené souhrny vybírají informace vztahující se k určitému tématu.
  - Aktualizační souhrny, zohledňující apriorní znalosti uživatele.
  - Uživatelsky zaměřené souhrny obsahují pouze informace týkající se oblastí zájmu jednotlivého uživatele nebo skupiny uživatelů.
- Na základě rozsahu:
  - Jednodokumentové souhrny.
  - Vícedokumentové souhrny.
- Podle jazyka:
  - Multijazykové.
  - Monojazykové.
- Dle použitého principu:
  - Heuristické metody.
  - Statistické metody (např. Naive Bayes, která je „metodou s učitelem“).
  - Grafové metody (např. PageRank, která je „metodou bez učitele“).
  - Algebraické metody (např. LSA, která je „metodou bez učitele“).

Jednotlivé principy se mohou vzájemně prolínat a doplňovat. V dalších částech si proto představíme alespoň některé zástupce jednotlivých skupin.

3 Klasické sumarizační metody

3.1 Heuristické metody

První pokusy s automatickou sumarizací jsou známé již z poloviny minulého století. Pracovaly extraktivním způsobem s využitím povrchových indikátorů pro výběr částí textu do výsledného extraktu. Za nejstarší je považován již zmíněný algoritmus publikovaný v [19]. Byl založen na předpokladu, že důležité termy se v textu často opakují, takže jejich frekvenci lze použít jako kriterium pro výběr vět do extraktu. Algoritmus nejprve zjistil počet výskytů jednotlivých slov (termů). Poté ohodnotil věty podle počtu a zjištěné významnosti v nich obsažených slov a do souhrnu pak zařadil věty s nejvyšším ohodnocením. Běžná slova (tzv. stop slova) nebyla do ohodnocování zahrnuta.

Jiné heuristické kriterium bylo použito v [6]. Využívalo skutečnosti, že důležitá slova se vyskytují v nadpisu, na začátku či na konci textu nebo bývají zdůrazněna přívlastky jako „významný“, „výsledný“, „důsledek“ apod. Kombinace pozičního kriteria spolu se zvýrazňujícím kontextem pak byla použita k ohodnocení a k výběru významných slov a jejich přítomnost ve větách indikovala vhodnost vět k zařazení do souhrnu.

3.2 Statistické metody

Důležitost termů z dokumentu se odráží ve frekvenci jejich výskytu. Této skutečnosti využíval již Luhnuv sumarizátor a je dobře známa z indexovacích mechanismů vyhledávacích systémů. Pokud se některé slovo ale bude vyskytovat v textech příliš často, jeho důležitost klesá. Proto je významnost termu *t* v dokumentu vyjadřována jako součin jeho frekvence výskytu *tf* a reciproční hodnoty počtu jednotek s jeho přítomností (inverted document frequency) *idf*. Do souhrnu jsou pak zařazovány věty, které jsou významné proto, že obsahují důležité termy. Postup sumarizace dokumentu může být popsán v následujících bodech:

1. Zkonstruuj pro každou větu *i* zpracovávaného dokumentu vektor frekvence termů tf_i .
2. Zkonstruuj vektor *D* inverzní frekvence termů v celém dokumentu.
3. Vypočti významnost každé z vět dokumentu pomocí skalárního součinu $tf_i \bullet D$.
4. Do výsledku zařaď věty s nejvyšším skóre.

Takto konstruovaný souhrn by pravděpodobně měl nedostatek. Příliš by akcentoval jedno hlavní téma dokumentu, které by bylo ve výsledku zastoupeno redundantně, proto bod 4 změníme, zařadíme do výsledku jen jednu větu *v* (s nejvyšším skóre) a dále provedeme:

5. Je-li délka výsledného souhrnu postačující, tak ukonči výpočet, jinak pokračuj dalším krokem.
6. Všechny termy, obsažené ve větě *v*, odstraň z vět dokumentu. Tím je z dokumentu současně odstraněna i věta *v*.
7. Opakuj výpočet od bodu 1.

V [10] byl použit obdobný postup se zdokonaleným vážením významnosti termů. Pomocí tezauru WordNet bylo prosté načítání frekvence nahrazeno „relevancí“. Čítač výskytu termu byl inkrementován i v případech nalezení výskytu synonym, hyponym (jedle pro strom), meronym (větev pro strom), či holonym (strom pro větev).

Důmyslnější statistickou metodu, která je založena na Bayesově klasifikačním vzorci poprvé použili v [17]. Věty z dokumentu je možné klasifikovat do dvou tříd: 1. zařazené do souhrnu a 2. nezařazené do souhrnu. K natrénování metody je potřebný korpus dvojic (originální texty a jim příslušné souhrny). Dále je třeba určit příznaky, na jejichž základě je prováděna klasifikace vět. Použité příznaky zahrnují přítomnost důležitých slov zjištěných na základě jejich frekvence, slova začínající velkým písmenem, délka věty, fráze se zdůrazňujícím slovem, pozici. Sumarizátor může určit pro každou větu dokumentu její pravděpodobnost zařazení nebo nezařazení do souhrnu na základě hodnot jejích příznaků a znalosti priorních pravděpodobností. Zjistí maximální aposteriorní pravděpodobnost, tj. nejpravděpodobnější hypotézu $h \in \{\text{zařadit, nezařadit}\}$ při daných hodnotách příznaků $f_1, f_2, \ldots f_k$.

$$P(h \mid f_1,f_2, \ldots f_k) = P(f_1,f_2, \ldots f_k \mid h) * P(h) / P(f_1,f_2, \ldots f_k) \qquad (1)$$

Obecnou Bayesovu formuli (1) lze za předpokladu nezávislosti příznaků (což sice není pravda, ale běžně se to toleruje) zjednodušit, použít místo složené pravděpodobnosti součin pravděpodobností a počítat pravděpodobnost zařazení dle vzorce pro naive Bayes klasifikátor.

$$P(h \mid f_1,f_2, \ldots f_k) = \prod P(f_j \mid h) * P(h) / \prod P(f_j) \qquad (2)$$

$P(f_j \mid h)$ představuje pravděpodobnost s jakou budou pro h = zařazení v souhrnu zařazeny věty s příznakem s hodnotou f_j, analogicky pro h = nezařazení se jedná o pravděpodobnost s jakou budou pro h = nezařazení do souhrnu nezařazeny věty s příznakem majícím hodnotu f_j. $P(f_j|h)$ se zjistí z trénovacího korpusu dvojic (text a souhrn). $P(f_j)$ představuje pravděpodobnost výskytu hodnoty příznaku f_j ve větách textového korpusu. P(h) je poměr počtu vět v souhrnu k celkovému počtu vět v trénovacím korpusu pro případ kdy h=zařazení a obdobně pro h = nezařazení. Do výsledného souhrnu se vloží potřebný počet vět, řazených podle spočtené pravděpodobnosti zařazení. Pro zamezení podtečení je vhodné vzorec převést do logaritmického tvaru a počítat zařazení vět na základě vyhodnocení:

Vyber takové $h \in \{\text{zařadit, nezařadit}\}$, pro které je větší hodnota $(\log P(h) + \sum \log P(f_j \mid h))$

Systém popsaný v [17] se poněkud liší od výše popsaného klasifikačního postupu. Neklasifikuje každou větu textu. Místo toho počítá pro každou větu skóre dané pravděpodobností jejího zařazení do souhrnu. Nejvýše hodnocených *n* vět pak tvoří souhrn. Výpočet se tím zrychlí cca dvojnásobně. Výsledky budou zhruba stejné, pokud použijeme převážně příznaky pozitivní pro zařazování do souhrnu.

4 Pokročilé sumarizační metody

4.1 Metody využívající souvislosti v textu

Do této skupiny můžeme zařadit jak metody využívající rétorické struktury textu, tak i metody pracující s anaforickými vztahy mezi větami. Společná je pro ně potřeba zvládnutí hlubších lingvistických znalostí než tomu je v případě již zmíněných statistických metod, které sumarizaci v podstatě převádí na klasifikaci a nebo algebraických metod, o kterých pojednáme později.

Teorie rétorických struktur (RST) zkoumá způsoby uspořádání projevu. Prostřednictvím rétorických relací zachycuje vzájemné vazby mezi jednotlivými částmi projevu (textu). Rozlišuje část zvanou nukleus, která obsahuje nejpodstatnější, tj. ústřední část textu s hlavními údaji a s ní svázané méně důležité, tzv. satelitní části. Nukleové a satelitní části jsou společně označovány jako textové jednotky a představují části vět nebo i celé věty. Rétorická struktura má podobu stromového grafu, jehož uzly jsou ohodnocovány podle jejich rétorické role. Uzel, který v jedné úrovni má vlastnost satelitu, může v nižší úrovni RS-stromu být nukleus a vázat se s dalším satelitním uzlem. V [20] popisovaný sumarizační program pracuje s RS-stromem, který je vygenerován rétorickým analyzátorem. Do souhrnu vybírá textové jednotky podle výše jejich umístění v RS-stromu. Pro krátké souhrny jsou vybírány pouze významné jednotky sdružené s vnitřními uzly stromu, které se nachází blízko jeho kořene. Čím delší souhrn je generován, tím od kořene vzdálenější významné jednotky textu jsou do něj zahrnuty. Princip metody je tedy založen na předpokladu, že RS-strom reprezentuje parciální uspořádání částí textu podle jejich důležitosti.

Extraktivní způsob sumarizace dává málo uspokojivé výsledky, pokud originál obsahuje časté anaforické výrazy. Anaforickým výrazem je slovo nebo fráze, odkazující zpět na nějaké dříve uvedené slovo nebo frázi (typickým příkladem jsou zájmena „ten“, „on“, …). Pro porozumění anaforickému výrazu je třeba znát jeho předchůdce. Je-li do extraktu vybrána věta obsahující anaforickou vazbu bez jejího kontextu, souhrn bude těžko srozumitelný. Soudržné vlastnosti textu jsou tvořeny relacemi mezi výrazy a byly také využity pro sumarizaci.

V [1] je uvedena metoda nazývaná „Lexikální řetězce“. Ve zpracovávaném textu nejprve vyhledá řetězce „příbuzných slov“. Příbuznými termy jsou takové, které jsou synonyma, hyperonyma, hyponyma, antonyma apod. K posouzení příbuznosti je využíván tezaurus Wordnet. Po zkonstruování těchto lexikálních řetězců, sumarizátor vypočte jejich skóre. To je určeno typem relací a jejich počtem v řetězci. Na základě hodnocení řetězců a jejich incidence s větami jsou pak hodnoceny samotné věty. Do souhrnu jsou vybírány ty věty, ve kterých se koncentrují řetězce s nejvyšším skóre.

V [3] je uveden podobný princip. Namísto lexikálních řetězců používá k ohodnocení vět tzv. objekty a jejich vazby. Objektem může být jak slovo, tak fráze nebo její variantní, či redukovaná forma. Vazby jsou dány odkazy mezi objekty. Věty jsou ohodnoceny na

základě počtu a míry referencí objektů v nich obsažených. Do souhrnu jsou vybrány věty, které obsahují často zmiňované objekty.

4.2 Metody modifikující původní text

Automatický extraktor není schopen (co se kvality výsledku týká) konkurovat ručně vytvořenému abstraktu. Počítač je sice schopný poměrně dobře rozpoznat klíčová témata v dokumentu, vytržení vět nebo odstavců originálu a jejich složení do souhrnu však téměř vždy naruší kontinuitu výsledného textu. Pokusy o vytvoření automatického sumarizátoru, který by pracoval neextraktivním způsobem a lépe zachoval souvislost textu se začaly objevovat před cca 10 ti lety. Lze je rozdělit do dvou skupin:

Prvá skupina místo překopírování celých vytipovaných vět nebo odstavců, konstruuje souhrn za pomoci jejich komprese. Sumarizátory pracující tímto způsobem jsou popsány např. v [13], [16], [27]. Vychází z předpokladu, že věty navržené do souhrnu bývají většinou dlouhé. Dlouhé věty totiž s větší pravděpodobností obsahují důležité termy. Často však také obsahují i méně důležité části. Úsek textu vybraný ke komprimaci je zpracován syntaktickým analyzátorem, který identifikuje v souvětích vedlejší věty, tj. kandidáty na vypuštění. Vyhodnocení vhodnosti či nevhodnosti kandidáta používá obvykle více hledisek. Např. počet zbylých důležitých termů ve zkrácené větě, hloubka vedlejší věty v syntaktickém stromu, počet odstraněných listů stromu, počet odstraněných vlastních jmen, porušení anaforických vazeb.

Do druhé skupiny lze zařadit [14], [21]. Generují věty nově, s pomocí „cut and paste“ operací. Operace mohou mít podobu:

- redukce věty - odstraňují irelevantní fráze, slova, vedlejší věty. Odstraněno může být ve výsledku i více komponent, pokud jsou shledány nezávažné.
- Větné kombinace - slučují texty z více vět. Obvykle jsou použity společně s redukcí slučovaných vět.
- Syntaktické transformace – přemístění větných částí na základě syntaktického rozboru. Doplňují větné redukce a kombinace.
- Parafrázování – nahrazuje fráze jejich parafrázemi (volné vyjádření obsahu jinými slovy).
- Generalizace / specifikace – nahrazuje fráze nebo vedlejší věty jejich obecnějším/specifičtějším popisem.
- Přeuspořádání – mění pořadí extrahovaných vět.

Zmiňovaný „cut and paste“ sumarizátor k realizaci výše uvedených operací využívá spolupráce se syntaktickým analyzátorem, s co-referenčním systémem, tezaurem a s rozsáhlým slovníkem.

4.3 Grafové metody

Iterační metody, o kterých v této části pojednáme, vznikly jako prostředek pro ohodnocování významnosti uzlů hypertextové struktury Webu. Všeobecnou známost si získaly algoritmy HITS[15] a PageRank [4]. S úspěchem byly použity i pro vyhodnocování autoritativnosti uzlů v sociálních sítích, zejména v jedné z jejich konkrétních podob, v citačních sítích [26], [8]. Sympatickou vlastností těchto metod pro vyhodnocování grafových struktur je jejich jazyková nezávislost a nepotřebnost hlubších lingvistických znalostí při jejich nasazení k sumarizaci.

Původně byly tyto úlohy aplikovány na orientované grafy. Nechť $G = (V, E)$ je orientovaný graf, s množinou vrcholů V a s množinou hran E, kde E je podmnožinou VxV. Pro daný vrchol V_i nechť $In(V_i)$ je množinou vrcholů, ze kterých vede větev do V_i a $Out(V_i)$ nechť je množina vrcholů do nichž vede větev z V_i.

Snad nejpopulárnějším algoritmem pro vyhodnocování významnosti vrcholů v grafu (ranking algorithm) je PageRank, používaný v Google k analýze Webu. Na rozdíl od jiných hodnotících metod (např. HITS) PageRank integruje do jediné formule vliv vstupních i výstupních charakteristik vrcholů. Pro každý vrchol tedy určuje pouze jediné PR (PageRank) skóre, dané vzorcem (3),

$$PR(V_i) = (1-d)/N + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|}, \quad (3)$$

ve kterém N je počet vrcholů a *d* je parametr (faktor tlumení) s hodnotou z intervalu 0 až 1. Je patrné, že PageRank vrcholu závisí na PageRanku ostatních vrcholů. Vzhledem k cykličnosti grafu je výpočet iteračním procesem, při kterém se propojené vrcholy navzájem ovlivňují.

Faktor *(1-d)* představuje pravděpodobnost, se kterou bude při procházení grafem proveden přechod na libovolný vrchol grafu. Naproti tomu *d* představuje pravděpodobnost přechodu podle větve vedoucí z vrcholu. Hodnotu *d* se doporučuje volit cca 0.8. Na počátečních hodnotách vrcholů nezáleží, volí se všechny stejné, se součtem 1. Výpočet konverguje během několika málo iterací.

Při použití iteračních vyhodnocování pro extraktivní sumarizaci reprezentují vrcholy grafu jednotlivé věty dokumentu. Větve grafu vyjadřují vazby mezi větami. Nejsou orientované, což není překážkou, algoritmus pracuje i s neorientovanými grafy. V tomto případě $In(V_i) = Out(V_i)$ tj. větve jsou považovány za vstupní i výstupní zároveň. Zatímco v případě sociálních sítí bývají větve grafu neohodnocené, při výběru vět v extraktivní sumarizaci je možné ohodnocením větví vyjádřit míru svázanosti věty V_i a V_j jako váhu w_{ij}. Originální vzorec pro PageRank nezahrnuje vážení větví. Proto [22] v systému TextRank formuli zmodifikovali na tvar (4):

$$PR(V_i) = (1-d)/N + d * \sum_{V_j \in In(V_i)} w_{ji} \frac{PR(V_j)}{\sum_{V_k \in Out(V_j)} w_{jk}} \quad (4)$$

Důležitou fází sumarizačního procesu v TexRanku je konstrukce grafu vazeb vět dokumentu. Pro určení a ohodnocení větví je zavedena relace podobnosti vět, která má význam překrytí kontextu. Lze ji chápat jako určité doporučení čtenáři, který čte větu V_x, aby si přečetl větu V_y, která pojednává o stejném konceptu. Doporučuje se určit váhu na základě počtu společných symbolů v obou větách, společných slov určité syntaktické kategorie, normalizovat váhy vzhledem k délce vět a tím předejít preferenci vět dlouhých. Formálně TextRank popisuje podobnost vět V_i a V_j, z nichž každá je reprezentovaná množinou N_i slov (resp N_j slov) W^i_1, W^i_2, ... W^i_{Ni} (resp. W^j_1, W^j_2, ... W^j_{Nj}) vzorcem (5).

$$Podobnost(V_i, V_j) = \frac{|\{W_k; W_k \in V_i \,\&\, W_k \in V_j\}|}{\log(|V_i|) + \log(|V_j|)} \quad (5)$$

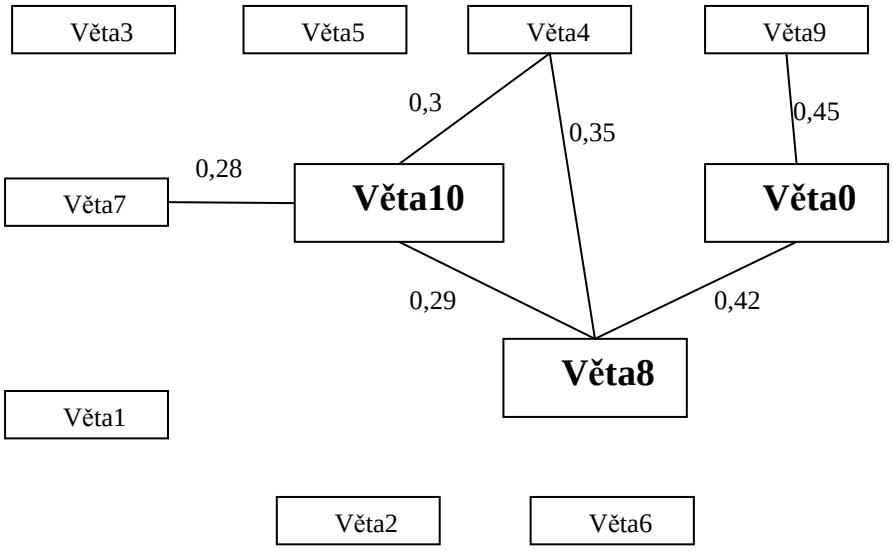
Podobnost lze určit i jinými způsoby. Po zkonstruování grafu podobnosti je použita formule (4). Výpočet je ukončen, když změny hodnot vrcholů jsou menší než zvolená mez. Věty s nejvyšším ohodnocením jsou pak vybrány do souhrnu.

Obdobný systém pro výpočet důležitosti vět je LexRank [7]. Podobnosti vět jsou zachyceny maticí, v níž hodnota prvků je dána kosinovou podobností (viz 7.1) příslušných vět. Podobnost je závislá na počtu překryvů slov. Dvě identické věty msjí podobnost 1, zatímco dvě věty se zcela odlišnými slovy mají podobnost 0. Demo verze je na adrese:

http://tangra.si.umich.edu/clair/lexrank.

Pro názornost uveďme konkrétní příklad originálního textu, jemu odpovídající graf podobnosti vět, ohodnocení vět a výsledný sumarizovaný text při nastavení cosinového filtru podobnosti na 25% a významnosti vybraných vět na 15%.

Původní text:
Každý už ví, že dovolenou je nutno kupovat jen u CK pojištěné proti úpadku. Ale kterou CK vybrat. Kam jít koupit svou vysněnou dovolenou. Možností je hodně.
I já jsem zpočátku obíhala cestovní kanceláře ve městě. Nyní ale využívám mnohem rychlejší a pohodlnější způsob. Vybírám si dovolenou na internetových stránkách. Jsou zde zájezdy všech velkých cestovních kanceláří a více než sta dalších ck. Do celého světa a za stejnou cenu jako u cestovní kanceláře. Navíc dostávám dárek - pojištění stornopoplatků v hodnotě 600Kč zdarma. To vše rychle a z pohodlí domova - internetem.



Obr. 1: Graf podobnosti vět při prahové hodnotě cosinu 0,25.

Pořadová čísla vět, výsledná ohodnocení a texty vět jsou uvedeny v Tab 1.

Číslo věty	**Ohodnocení věty**	**Text věty**
9	0.08699246309	Navíc dostávám dárek - pojištění stornopoplatků v hodnotě 600Kč zdarma.
7	0.08024629241	Jsou zde zájezdy všech velkých cestovních kanceláří a více než sta dalších ck.
5	0.02173913043	Nyní ale využívám mnohem rychlejší a pohodlnější způsob.
2	0.02173913043	Kam jít koupit svou vysněnou dovolenou.
10	0.20649587529	To vše rychle a z pohodlí domova - internetem.
8	0.20419247759	Do celého světa a za stejnou cenu jako u cestovní kanceláře.
6	0.02173913043	Vybírám si dovolenou na internetových stránkách.
3	0.02173913043	Možností je hodně.
0	0.15353727589	Každý už ví, že dovolenou je nutno kupovat jen u CK pojištěné proti úpadku.
4	0.13810083309	I já jsem zpočátku obíhala cestovní kanceláře ve městě.
1	0.02173913043	Ale kterou CK vybrat.

Tab.1: Ohodnocení významnosti vět sumarizátorem LexRank

Výsledný souhrn :
Každý už ví, že dovolenou je nutno kupovat jen u CK pojištěné proti úpadku. Do celého světa a za stejnou cenu jako u cestovní kanceláře. To vše rychle a z pohodlí domova - internetem.

5 Latentní sémantická analýza a sumarizace

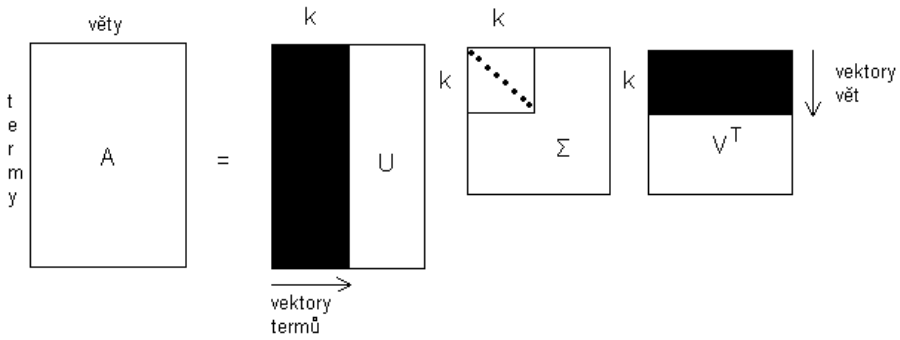
5.1 Princip latentní sémantické analýzy

Latentní sémantická analýza (LSA) je algebraická technika, dovolující automaticky analyzovat vztahy mezi termy a dokumenty, či termy a větami. Používá metodu rozkladu matic singulární dekompozicí (SVD). SVD je numerický proces, který se používá při redukci dat. Byly navrženy algoritmy, které singulární dekompozicí řeší klasifikaci nebo vyhledávání dokumentů (latentní sémantické indexování). SVD byla poprvé použita pro sumarizaci v [9] a zdokonalena v [28]. Námi navržený princip popíšeme nejprve pro sumarizaci jednoho dokumentu. Modifikace pro složitější úlohy uvedeme v další kapitole.

Proces začíná vytvořením matice termů proti větám $A = [A_1, A_2, \ldots, A_n]$, kde každý sloupcový vektor A_i reprezentuje vektor frekvencí termů ve větě i dekomponovaného dokumentu. Pokud dokument obsahuje m termů a n vět, získá se matice A o rozměrech $m \times n$. Matice A je zpravidla řídká, protože normálně se každé slovo v každé větě nevyskytuje. Singulární dekompozice matice A je potom definována jako:

$$A = U\Sigma V^T, \tag{6}$$

kde $U = [u_{ij}]$ je $m \times n$ sloupcově ortonormální matice, jejíž sloupce se nazývají levé singulární vektory, $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ je $n \times n$ diagonální matice, jejíž diagonální prvky jsou nezáporná singulární čísla seřazená sestupně a $V = [v_{ij}]$ je $n \times n$ ortonormální matice, jejíž sloupce se nazývají pravé singulární vektory (viz obr. 1). Rozměr matic je redukován na k dimenzí, kde $k < n$, takže U je redukována na $m \times k$, Σ na k×k a V^T $k \times n$.



Obr 2.: Singulární dekompozice matice A

Na SVD rozklad matice *A* termů proti větám se můžeme dívat ze dvou pohledů. Z matematického pohledu SVD zprostředkovává mapování mezi *m*-dimenzionálním prostorem vektorů frekvencí termů a *k*-dimensionálním singulárním vektorovým prostorem. Ze sémantického pohledu SVD poskytuje latentní sémantickou strukturu dokumentu reprezentovaného maticí *A*. Tato operace vyjadřuje rozklad originálního dokumentu do *k* lineárně nezávislých bázových vektorů reprezentujících hlavní témata textu. Každý term i věta dokumentu jsou indexovány těmito bázovými vektory. Unikátní vlastností singulárního rozkladu je schopnost zachytit a modelovat vnitřní vztahy mezi termy tak, že může sémanticky shlukovat termy a věty. Dále, jak je demonstrováno v [2], pokud se v dokumentu často vyskytuje určitá kombinace slov, pak bude tato kombinace zachycena a reprezentována jedním ze singulárních vektorů. Velikost odpovídajícího singulárního čísla indikuje významnost kombinace v dokumentu. Každá věta obsahující tuto kombinaci slov bude promítnuta podél odpovídajícího singulárního vektoru a věta, která nejlépe reprezentuje tuto kombinaci, bude mít největší hodnotu v tomto vektoru. Každá kombinace slov popisuje určité téma dokumentu. Lze tedy na základě předchozích faktů říci, že každý singulární vektor reprezentuje určité téma dokumentu a velikost korespondujícího singulárního čísla reprezentuje významnost tohoto tématu [9]. Shrneme-li předchozí výklad, tak matice *A* mapuje termy do jednotlivých vět, redukovaná matice *U* mapuje termy do *k* nejvýznamnějších témat a redukovaná matice *V* mapuje věty do *k* nejvýznamnějších témat.

5.2 Použití LSA pro sumarizaci

Na základě předchozí diskuse jsme navrhli sumarizační metodu. Tato metoda využívá singulární rozklad matice termů proti větám, konkrétně matici V^T, která popisuje míru významnosti vět v hlavních tématech dokumentu. Algoritmus navržený v [9] jednoduše vybírá pro každé téma nejvýznamnější větu tak, že postupně pro j = 1 až do potřebného počtu *P* vět souhrnu vybere j-tý pravý singulární vektor z V^T. Každá věta je reprezentována

sloupcovým vektorem $[v_{j1}, v_{j2}, \ldots, v_{jk}]^T$. Do souhrnu zařadí tu větu, která má největší indexovou hodnotu v j-tém pravém singulárním vektoru.

Nevýhodou takového postupu je stejná důležitost všech *P* v souhrnu obsažených témat. Jejich významnost se však může výrazně lišit, což lze identifikovat v matici Σ. Navrhli a ověřili jsme proto změnu kriteria výběru dovolující zařadit věty, jejichž vektorová reprezentace v matici $\Sigma^2 \times V^T$ má největší délku. Násobením Σ^2 zohledníme statistickou významnost hlavních témat, která je úměrná kvadrátu příslušného singulárního čísla, jak bylo dokázáno v [5]. Formálně vyjádřeno, počítáme v *k* rozměrném latentním prostoru témat délku vektoru s_r pro r-tou větu dle vzorce:

$$s_r = \sqrt{\sum_{i=1}^{k} v_{ri}^2 * \sigma_i^2} \tag{7}$$

V experimentech jsme dimenzi latentního prostoru omezili zvoleným procentem z celkového počtu dimenzí. Je možné použít i poklesu singulárních čísel na zlomek největšího. Do souhrnu je zařazován žádaný počet vět, jejichž hodnoty *s* jsou největší. Důležité téma tak může být v souhrnu zastoupeno více větami. LSA byla pro sumarizaci použita i v dalších modifikacích. Např. po SVD rozkladu byla zpětně rekonstruována redukovaná matice A^R a na její „věty" pak aplikován výše uvedený grafový postup [23]. Jiný přístup zařazuje počet vět vztahujících se k tématu na základě procentuálního podílu příslušného singulárního čísla k součtu všech singulárních čísel [34].

SVD není jedinou algebraickou metodou, která se uplatňuje v úlohách zpracování textu. Jinou metodou s obdobnými schopnostmi je NMF (non-negative matrix factorization), která rozkládá matici *A* na dvě matice *W* a *H*. Jejich prvky rovněž reprezentují termy a věty v prostoru témat. Protože jsme chtěli využívat informaci o důležitosti témat z matice Σ, NMF jsme zatím nevěnovali výraznou pozornost.

6 Vícedokumentová sumarizace a nové sumarizační úlohy

Před zhruba sedmi lety se pozornost týmů zabývajících se sumarizací začala soustřeďovat na vícedokumentovou sumarizaci a s ní související úlohy jako je sumarizace aktualizační (update) [12], cílená (focuced), kontrastová (contrastive) [33], či mínění (sentiment). Vícedokumentová sumarizace oproti jednodokumentové zavádí nový problém – je třeba zabránit zařazení do souhrnu vět z různých dokumentů, ale se stejným obsahem.

V prvé fázi zpracování postupujeme stejně jako při sumarizaci jednoho dokumentu, pracujeme však se všemi větami množiny dokumentů. Některou z dříve uvedených metod ohodnotíme věty skórem vhodnosti jejich zařazení do souhrnu.

Ve druhé fázi vybíráme sestupně podle skóre jednotlivé věty. Před jejich zařazením do souhrnu ale navíc ověřujeme, zda v souhrnu již není podobná věta. Podobnost je možné měřit např. kosinem úhlu mezi větami ve vektorovém prostoru termů množiny dokumentů. Pro verdikt o zařazení/nezařazení je třeba zvolit prahovou hodnotu kosinu. Volba prahu závisí na rozložení hodnot skóre vět, takže určením prahu musíme nastavit rozumný poměr mezi podobností a skórem vět souhrnu. Skóre vět v sobě odráží počet zvažovaných témat množiny dokumentů. Proto volba prahu se může lišit podle zpracovávané oblasti a je vhodné ji experimentálně ověřit. Jinou možností je použít iterační formuli (10) z odst. 6.1.

Problém, který se projevil při našich experimentech, bylo upřednostňování delších vět. Přirozeně, dlouhé věty obsahují více významných termů. Skóre vět bylo proto děleno koeficientem, jehož velikost závisela na délce věty. Vyhovující výsledky byly dosaženy již při poměrně malém počtu témat, cca do 10. Byla ovšem zohledněna jejich významnost násobením V^T mocninou Σ [30].

6.1 Aktualizační sumarizace

V případě aktualizační sumarizace předpokládáme, že uživatel má z dané oblasti předchozí znalosti, které získal přečtením množiny dokumentů C_{old}. Dále máme množinu dokumentů C_{new}, které dosud nečetl a chce se seznámit s jejich souhrnem. Do souhrnu však nechceme zařazovat ty informace z C_{new}, které již byly obsaženy v C_{old}. Předpokládáme tedy čtenáře s dokonalou pamětí.

Popišme řešení pomocí LSA modelu [29]. Aplikujeme SVD odděleně na matice A_{new} a A_{old} vytvořené z C_{new} a C_{old}. Získáme redukované matice U_{new} a U_{old}, jejichž sloupce obsahují témata množin dokumentů, vyjádřená v lineárních kombinacích původních termů. Pro každé *nové* téma, dané sloupcem matice U_{new}, (označme index tohoto sloupce t), vyhledáme nejpodobnější *staré* téma dané sloupcem matice U_{old}. Kosinová podobnost těchto dvou vektorů udává míru redundance nového tématu *red(t)*.

$$red(t) = \max_{i=1}^{k} \frac{\sum_{j=1}^{m} U_{old}[j,i] * U_{new}[j,t]}{\sqrt{\sum_{j=1}^{m} U_{old}[j,i]^2} * \sqrt{\sum_{j=1}^{m} U_{new}[j,t]^2}}, \tag{8}$$

kde k je počet sloupců matice U_{old}, tj. počet hlavních témat v redukovaném latentním prostoru. Novost tématu t počítáme vztahem $1 - red(t)$, a protože důležitost tématu je obsažena v odpovídajícím singulárním čísle $\sigma(t)$, počítáme aktualizační skóre $us(t)$ tématu t dle vzorce:

$$us(t) = \sigma(t) * (1 - red(t)) \tag{9}$$

Z vypočtených skóre sestavíme diagonální matici US a vynásobením $US \cdot V^T$ dostaneme tak matici F, která v sobě agreguje novost i důležitost nových témat. Následuje zařazování vět do aktualizačního souhrnu. První je věta, která má nejdelší vektor v matici F. Označme jej f_{best}. Informaci, kterou jsme touto větou začlenili do souhrnu, je třeba odečíst od ostatních vektorů (vět) f. Přepočítáme proto sloupce matice F.

$$F_{i+1} = F_i - \frac{f_{best} \cdot f_{best}^T}{|f_{best}|^2} \cdot F_i \tag{10}$$

Proces zařazování do souhrnu probíhá iteračně, až do získání potřebné délky souhrnu.

6.2 Další aktuální sumarizační úlohy

Stručně a bez nároku na úplnost výčtu se v tomto odstavci zmíníme o sumarizačních úlohách, které stejně jako aktualizační byly motivovány sumarizací více dokumentů.

Kontrastová sumarizace provádí analýzu dokumentů s cílem nalézt rozdíly v jednotlivých dokumentech. Výsledkem je nejen souhrn společný všem dokumentům, ale i informace o důležitých tématech specifických pro jednotlivé dokumenty. Zkoumání rozdílnosti dokumentů přes její praktickou využitelnost bylo věnováno velmi málo pozornosti v porovnání se zkoumáním jejich podobnosti. Dosud jsme nenalezli práci, která by tuto úlohu řešila metodou LSA. Nabízí se přitom možnost po provedení vícedokumentové sumarizace provést sumarizaci jednotlivých dokumentů a např. kosinovou mírou porovnat rozdílnost jejich témat s tématy celkového souhrnu. Překročí-li rozdíl zvolenou mez, pak zařadit příslušné věty do rozdílových souhrnů obdobným postupem jako byl popsán výše.

Sumarizace mínění zpracovává množinu dokumentů *D*, které obsahují hodnocení nějaké entity (zboží, služeb apod). Výsledkem je souhrn *S*, který reprezentuje průměrný názor o této entitě. Pracuje s polarizační funkcí, která zobrazuje části textů (fráze, věty) do číselných hodnot, odlišujících kladný a záporný názor. Tato funkce je realizována speciálními lexikony. Polarizované části textů jsou načítány, zprůměrovány a výsledné skóre určuje jemu odpovídající části textu, které jsou zařazeny do souhrnu.

Cílená sumarizace zahrnuje do vstupních dat i uživatelem specifikovanou informaci. Ta může být zadána formou dotazu, nebo tématem o které se zajímá. Množinou sumarizovaných dokumentů bývají v tomto případě často webové stránky. Základem je opět vícedokumentová sumarizace, do výsledku jsou však přednostně zařazovány věty, jejichž téma odpovídá informaci od uživatele. K tomu je nutné zavést metriku témat porovnávající téma dotazu nebo klíčových slov s tématy vět. Řešení úlohy tohoto typu jsme s použitím LSA popsali v [32]. Summarizer of Web Topics (SWEeT) je volně přístupný na

http://tmrg.kiv.zcu.cz:8080/sweet.

Odpovídá na anglické a české dotazy, ze kterých extrahuje významné termy. Ty pak použije vyhledávací modul k prohledání předdefinovaných domén vyhledávači Google a Seznam. Prvých 10 dokumentů je předáno analyzátoru, který vybere z HTML struktury vlastní texty a předá je v XML podobě extrakčnímu modulu. Dále se provádí LSA extrakce vět. Proti dříve popsanému s tím rozdílem, že termům z dotazu je přiřazena větší váha v matici *A*. Následuje komprese vět, jejich uspořádání, korekce entit a posléze zobrazení výsledku uživateli.

Výsledek na dotaz: kdo vyhraje komunální volby v Praze

Téměř všechny politické strany v Praze už kandidátky schválily, změny mohou nicméně dělat až do 10. srpna. Komunální volby se rozhodnutím prezidenta konají 15. a 16. října. Na podobné předpovědi je však brzo. Jisté je, že TOP 09 zatím dělá vše pro to, aby komunální volby v Praze vyhrála. A ODS vše pro to, aby je prohrála.

Použité zdroje:

- Jak fungují nábory v ČSSD? Kdo přivede víc „černých duší“, vyhraje
- KOMENTÁŘ: TOP Tůma. Lidé chtějí osobnosti, tady jedna je
- Podpořím Nečase jako šéfa ODS i premiéra, odpověděl čtenářům Bendl
- Analytik řekl online, kdo by mohl vyhrát volby
- ODS nachystala v Praze past, do níž může sama spadnout …

Obr.3. Příklad výstupu systému SWEeT

7 Vyhodnocování kvality sumarizace

Způsoby vyhodnocování kvality souhrnu jsou podrobněji popsány v [31]. Kromě ručního, subjektivního ohodnocení souhrnu anotátorem, existují i automatické vyhodnocovací metody. Míry a metody, které jsou používané k vyhodnocení, mohou být rozděleny do dvou, dále se podrobněji větvících skupin:

- Přímé (intrinsic), posuzují kvalitu na základě:
  - Porovnání lingvistické kvality textu, která může zohledňovat:
    - Gramatickou správnost,
    - Neredundantnost,
    - Srozumitelnost,

        - Strukturu a souvislost.
    - Porovnání obsahu textu, s ideálním souhrnem, k čemuž může být použito:
        - Ko-selekčních přístupů pracujících s pojmy (Přesnost, Úplnost, F-score, či Relativní užitečnost),
        - Podobnostních měr (kosinová podobnost, nejdelší společný podřetězec, společné n-gramy (Rouge), překrytí obsahu, ohodnocování vět (Pyramids),

- Nepřímé (extrinsic), posuzují kvalitu způsobem, jak se souhrn uplatňuje v určité úloze. K ohodnocení je možné použít:
    - Metody pro kategorizaci dokumentů,
    - Metody pro vyhledávání informací,
    - Metody pro zodpovídání dotazů.

Některé z pojmů jsou dostatečně vysvětlující, o těch se v následujícím komentáři zmíníme jen stručně, nebo je pomineme.

7.1 Přímé způsoby hodnocení kvality

K lingvistické kvalitě není příliš co doplňovat. Snad jen upozornit na nebezpečí zhoršení srozumitelnosti v případě vypuštění vět s podstatnými jmény a jejich zastoupení zájmennými vazbami v souhrnu. Problém anaforických vztahů, které vznikají v souhrnu, není ještě uspokojivě vyřešen. Pokus o možné řešení je popsán v [33]. Lingvistická kriteria nejsou vesměs dosud automaticky vyhodnotitelná. Anotátoři musí souhrny oznámkovat ručně.

Ko-selekční techniky používají míry známé z oblasti vyhledávání informací (IR information retrieval) a klasifikace. Nejznámějšími měrami jsou přesnost P, úplnost R (recall) a F-skóre. K vyhodnocení strojově vytvořeného souhrnu používají ideální (anotátorem vytvořený) souhrn. Přesnost je dána počtem vět, které se vyskytují současně v hodnoceném i v ideálním souhrnu, děleném počtem vět hodnoceného souhrnu. Úplnost je dána počtem vět, které se vyskytují současně v hodnoceném i v ideálním souhrnu, děleném počtem vět ideálního souhrnu. F-skóre je kombinovanou mírou, obvykle je vyhodnocováno formulí pro harmonický průměr P a R: F-skóre = (2 * P *R) / (P + R).

Relativní užitečnost RU eliminuje nedostatek výše uvedených ko-selekčních měr. Nedostatek spočívá ve striktním započítávání při výpočtu P, R, i F pouze vět z ideálního souhrnu. Hodnocení pomocí RU je proto založeno na přiřazení priority (určující pořadí začlenění do souhrnu) všem větám sumarizovaného textu. Ohodnocení vět prioritou je prováděno anotátory. Metrika, která udává kvalitu souhrnu je pak dána formulí sčítající bodový zisk vět obsažených v souhrnu.

Podobnostní míry mají rovněž svůj původ v oblasti IR. Oproti ko-selekčním technikám mohou rozpoznat věty s podobným obsahem a zohlednit tuto skutečnost při hodnocení. Tyto metody totiž počítají podobnost extraktů na nižší úrovni než jenom na úrovni celých vět. Jsou použitelné jak pro výpočet podobnosti vyhodnocovaného souhrnu s ideálním referenčním souhrnem, tak pro výpočet průměru z podobností vyhodnocovaného souhrnu s více manuálně připravenými souhrny, ale i pro vyhodnocení podobnosti s originálním dokumentem, tedy bez použití ideálního souhrnu. V dalším výkladu předpokládejme porovnávání podobnosti s originálem.

Nejpopulárnější podobnostní mírou je kosinová podobnost. Označíme-li X hodnocený souhrn a Y originální text, pak kosinová podobnost souhrnu s originálem je dána vzorcem:

$$\cos(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum_i (x_i)^2} * \sqrt{\sum_i (y_i)^2}} \tag{11}$$

Dokumenty *X* a *Y* jsou reprezentovány vektory v prostoru slov, obvykle s použitím tf-idf vah. Použití kosinu úhlu mezi vektory obou dokumentů současně eliminuje vliv jejich rozdílné délky.

Kosinová míra může být použita i v latentním prostoru témat namísto v prostoru slov. Ověření vhodnosti takového postupu jsme zveřejnili v [31]. Použití singulární dekompozice nám nabízí několik možných způsobů jak měřit podobnosti dokumentů.

Nejprostším způsobem je měření podobnosti hlavních témat originálu a souhrnu. Hlavní téma je skryto v prvním levém singulárním vektoru. Proto provedeme rozklad původního dokumentu a porovnávaného souhrnu, zjistíme jejich první levé singulární vektory a vypočteme podobnost jako kosinus úhlu podle vzorce:

$$\cos(\varphi) = \sum_i^m uo_i * us_i \tag{12}$$

ve kterém *uo* představuje prvý levý singulární vektor rozkladu originálu, *us* prvý levý singulární vektor rozkladu souhrnu a *m* je počet různých slov originálního textu.

Jistě není překvapením, že kromě měřítka podobnosti daného pouze hlavním tématem, lze hodnotit podobnost i z pohledu *n* hlavních témat porovnávaných dokumentů. Opět nejprve vytvoříme singulární rozklady obou dokumentů. Pak pro oba dokumenty vynásobíme matice *U* a matice kvadrátů singulárních čísel Σ^2. Získáme tím matice B_o (pro originální dokument) a B_s (pro souhrn):

$$B_o = U_o \cdot \Sigma_o^{\ 2}, \tag{13}$$

$$B_s = U_s \cdot \Sigma_s^{\ 2}. \tag{14}$$

Násobením zohledníme statistickou významnost hlavních témat, která je úměrná kvadrátu příslušného singulárního čísla [5]. Pro každý vektor termu (řádek matice B) pak spočítáme jeho délku. Výpočet provedeme jak pro souhrn, tak i pro referenční dokument podle vzorce:

$$d_k = \sqrt{\sum_{i=1}^{n} b_{ki}^2}, \tag{15}$$

kde d_k je délka vektoru *k*-tého termu (jeho důležitost v latentním prostoru), *n* je počet nejvýznamějších témat. Z délek vektorů termů sestavíme výsledný vektor délek termů v latentním prostoru vzniklém singulární dekompozicí. Získáme tím dva vektory. Jeden pro souhrn (*ds*) a druhý pro originální dokument (*do*). Tyto vektory potom znormalizujeme. Pro změření jejich podobnosti použijeme opět kosinovou míru:

$$\cos\varphi = \sum_{i=1}^{m} do_i \cdot ds_i \cdot \tag{16}$$

Tato metoda má výhodu oproti předchozí. Pokud bude originální dokument obsahovat dvě či více přibližně stejně důležitých témat (odpovídající singulární čísla budou mít přibližně stejnou hodnotu), pak se může stát, že v extraktu tato stejně důležitá témata budou neprávem potlačena. Tuto nevýhodu odstraníme, pokud hodnotíme podle více témat.

Jiné podobnostní míry vychází z počtu slov resp. lemmat společných oběma dokumentům, či z počtu slov jejich nejdelšího společného podřetězce a z počtu editačních úprav potřebných k jeho získání.

Populárním způsobem měření kvality na bázi podobnosti textu je ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [18]. Jedná se o automatickou metodu, v současnosti používanou i k vyhodnocování soutěží pořádaných konferencí TAC, dříve DUC. ROUGE pracuje s kolekcí měr, které jsou založeny na podobnosti n-gramů (tj. n po sobě následujících slov textu). Rouge-n skóre kandidátního souhrnu je vyhodnoceno podle vzorce:

$$ROUGE-n = \frac{\sum_{C \in RSS} \sum_{gram_n \in C} Pocet_{spolu}(gram_n)}{\sum_{C \in RSS} \sum_{gram_n \in C} Pocet(gram_n)}, \quad (17)$$

kde RSS je množina referenčních souhrnů vytvořených anotátory, $Pocet(gram_n)$ je počet n-gramů v referenčním souhrnu a $Pocet_{spolu}(gram_n)$ je maximální počet n-gramů společně se vyskytujících jak v hodnoceném, tak i v referenčním souhrnu. Další používaná ROUGE skóre jsou ROUGE-SU4, pracující s bigramy, ale dovolující vypustit až 4 unigramy z bigramových komponent, nebo ROUGE-L, které pracuje s nejdelší společnou subsekvencí.

Poslední ze způsobů přímého hodnocení, který zmíníme, se nazývá Pyramids [24]. Spočívá v určování tzv. *summarization content units* (SCU), kterými jsou věty nebo jejich části. SCU jsou určeny a ohodnoceny podle počtu jejich výskytu v *n* ručně vytvořených souhrnech. Vyskytují-li se ve více souhrnech, získají vyšší hodnocení. Vzniká tak pyramida, na jejímž vrcholu jsou nejlepší SCU. Pyramida je pak použita k obodování hodnocených souhrnů.

7.2 Nepřímé způsoby hodnocení kvality

Pro tuto skupinu je charakteristické, že k určení kvality používá míry, jakou se hodnocený souhrn uplatní v jiné úloze z oblasti textminingu. Zjišťují kvalitu použitím automatických souhrnů pro daný praktický úkol. Testovat je možné například zvýšení rychlosti či přesnosti vyhledávání dokumentů, pokud je vyhledávání založené na extraktech místo na plných dokumentech např. metodou *Relevance correlation (RC)*. Dalším možným měřením je úspěšnost kategorizace dokumentů do tématických skupin, pokud se indexují extrakty místo původních dokumentů.

Korelace relevance (*korelace důležitosti*) je technika, která umožňuje měřit relativní pokles výkonu získávání informací, pokud se indexují souhrny místo plných dokumentů [25]. Předpokládejme, že máme dotaz Q a kolekci D dokumentů D_i. Vyhledávací systém seřadí dokumenty D_i podle jejich relevance k dotazu Q. Potom provedeme substituci plných dokumentů za souhrny S_i a stejný vyhledávací systém seřadí dokumenty S_i podle jejich relevance k dotazu Q. Pokud jsou souhrny dobrou náhradou původních dokumentů, předpokládá se, že pořadí v obou případech budou podobná. Existuje několik metod pro měření podobnosti pořadí (Kendall´s tau, Spearman´s rand correlation). Protože však máme navíc k dispozici z vyhledávacího systému relevanci jednotlivých dokumentů k dotazu, můžeme spočítat *RC* následujícím způsobem:

$$RC = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}, \tag{18}$$

kde x_i je relevance dokumentu D_i k dotazu Q, y_i je relevance souhrnného dokumentu S_i k dotazu Q. $\bar{x}$ (resp. $\bar{y}$) je průměrná relevance dokumentů z D (resp. z S) k dotazu Q.

Jiná metoda zjišťuje vhodnost použití souhrnů místo plných textů pro kategorizaci [11]. Pro měření je potřebná zatříděná kolekce dokumentů. Při tomto způsobu testování se ke klasifikaci používá automatický klasifikátor. Z důvodu oddělení chyby klasifikátoru a chyby sumarizátoru je pak nutné použití některých základních hodnot pro porovnání. Výsledné hodnoty klasifikace extraktů jsou proto porovnávány např. s výsledky hodnocení původních dokumentů nebo hodnocení náhodně vybraných vět. Posledním problémem zůstává míra určující kvalitu extraktu. Obecně se používají koeficienty přesnosti kategorizace P a úplnosti kategorizace R, vyhodnocované dle (19):

$$P = \frac{p}{q}, \quad R = \frac{p}{r}, \tag{19}$$

kde p je počet tříd, do kterých je dokument správně zatříděn klasifikátorem, q je celkový počet tříd, do kterých je dokument klasifikátorem zařazen a r je počet relevantních tříd, do kterých byl dokument klasifikovaný při předchozím ručním zatřiďování. Potom P a R pro celou kolekci je průměrem P a R přes všechny dokumenty. Z definice je možné vidět, že oba ukazatele spolu souvisí a zvyšováním jednoho se druhý bude snižovat. Při zařazení dokumentu do co nejvyššího počtu tříd bude vysoká úplnost, při snižování počtu tříd se bude zvedat přesnost. Z toho důvodu se pak používá pro hodnocení klasifikace např. průměr z obou hodnot nebo již dříve zmíněné *F-skóre*.

8 Závěr

Článek popisuje vývoj a současný stav automatické sumarizace textu. Vzhledem k pokroku, který sumarizace zaznamenala v posledním desetiletí, jsme se věnovali zejména extraktivním způsobům, které dle našeho úsudku budou ještě dlouho dominantní formou strojového vytváření souhrnů. Abstraktivním způsobům bylo ve sledovaném období věnováno mnohem méně prací. Vyžadují buď ruční vytváření šablon, které jsou strojově doplňovány extraktivní technikou, nebo hlubší analýzu textu a systém pro generování přirozeného jazyka. Oba přístupy jsou doménově závislé a náročné na ruční zpracování. Kromě přehledu sumarizačních metod jsme se věnovali i způsobům vyhodnocování a měření kvality sumarizace. Určení kvality souhrnu považujeme za stejně důležitou úlohu jako je sumarizace sama. Zvýšenou pozornost jsme věnovali použití metody singulární dekompozice, která nás zaujala svou jazykovou nezávislostí a elegancí matematického aparátu. Na jejím použití v pokročilých sumarizačních úlohách a zdokonalení integrováním s dalšími metodami chceme dále pracovat.

Literatura

1. Barzilay, R., Elhadad, M.: Using Lexical Chains for Text Summarization. In: *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, (1997), 10–17.

2. Berry M.W., Dumais S.T., O'Brien G.W. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review*. 1995.Boguraev, B., Kennedy, C.: Salience-based content characterization of text documents. In: I. Mani and M.T. Maybury. (Eds.), *Advances in Automatic Text Summarization*, The MIT Press (1999), 111-120.
3. Boguraev, B., Kennedy, C.: Salience-based content characterization of text documents. In: *Advances in Automatic Text Summarization*, MIT Press (1999),99-110.
4. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. In: *Computer Networks and ISDN Systems*, 30, (1998), 1–7.
5. Ding, Ch.: A Probabilistic Model for Latent Semantic Indexing. In: *Journal of the American Society for Information Science and Technology*, 56(6), (2005), 597-608.
6. Edmundson, H.P.: New Methods in Automatic Extracting. In: *Journal of the Association for Computing Machinery* 16(2). (1969) 264–285.
7. Erkan, G., Radev, D., G.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. In: *Journal of Artificial Intelligence Research 22*.(2004),457-479
8. Fiala D., Rousselot F., Jezek K.: PageRank for Bibliographic Network. In: *Scientometrics*, 76(1), Springer (2008), 135-158.
9. Gong, X., Liu X.: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In: *Proceedings ACM SIGIR.* New Orleans, USA (2001), 19-25.
10. Hovy, E., Lin, C-Y.: Automated Text Summarization in SUMMARIST. In: I. Mani and M.T. Maybury (Eds.), *Advances in Automatic Text Summarization*, The MIT Press, (1999), 81–94.
11. Hynek, J., Ježek, K.: Practical Approach to Automatic Text Summarization. In: *Proceedings 7. Conf. ELPUB '03*. Guimaraes, Portugal (2003), 378-388.
12. Jezek, K., Steinberger, J.: Automatic Text Summarization (The state of the art and new challenges). In: *Proceedings of Znalosti 2008*, Bratislava, Slovakia, (2008), 1–12.
13. Jing, H.: Sentence Reduction for Automatic Text Summarization. In: *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, USA, (2000), 310–315.
14. Jing, H., McKeown, K.: Cut and Paste Based Text Summarization. In: *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, USA, (2000), 178–185.
15. Kleinberg, J.M.: Authoritative sources in a hyper-linked environment. In: *Journal of the ACM*, 46(5), (1999), 604-632.
16. Knight, K., Marcu, D.: Statistics-Based Summarization Step One: Sentence Compression. In: *Proceeding of The 17th National Conference of the American Association for Artificial Intelligence*, (2000), 703–710.
17. Kupiec, J., Pedersen, J.O., Chen, F.: A Trainable Document Summarizer. In: *Research and Development in Information Retrieval.* (1995) 68–73.
18. Lin, Ch.,: ROUGE: A Package for Automatic Evaluation of Summaries. In: *Proceedingsot the Workshop on Ewxt Summarization Branches Out*, Barcelona, Spain, (2004).
19. Luhn, H.P.: The Automatic Creation of Literature Abstracts. In: *IBM Journal of Research Development* 2(2). (1958) 159–165.
20. Marcu, D.: From Discourse Structures to Text Summaries. In: *Proceedings of the ACL97/EACL97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, (1997), 82–88.

21. McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., Eskin, E.: From Discourse Structures to Text Summaries. In: *Towards Multidocument Summarization by Reformulation: Progress and Prospects, AAAI/IAAI*, (1999), 453–460.
22. Mihalcea, R., Tarau, P.: Text-rank - bringing order into texts. In: *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, (2004),
23. Murray, G., Renals, S., Carletta J.: Extractive Summarization of Meeting Recordings. In: *Proceedings of Interspeech*, Lisboa, Portugal, (2005).
24. Nenkova, A., Passonneau, R.: Evaluating Content Selection in Summarization: The Pyramid Method. In: *Document Understanding Conference*, Vancouver, Can., (2005).
25. Radev R., at all, Evaluation Challenges in Large-Scale Document Summarization. In: *Proceedings ACL Conference*, Sapporo, Japan, (2003).
26. Sidiropoulos A., Manolopoulos Y. A Citation-Based System to Assist Prize Awarding. In: *SIGMOD Record*, 34(4), (2005), 54-60.
27. Sporleder, C., Lapata, M.: Discourse chunking and its application to sentence compression. In: *Proceedings of HLT/EMNLP*, Vancouver, Canada, (2005), 257–264.
28. Steinberger J., Ježek K. Text Summarization and Singular Value Decomposition. In: *ADVIS '04*. Izmir, Turky, LNCS 3261, (2005), 245-254.
29. Steinberger, J., Ježek, K.: Update Summarization Based on Latent Semantic Analysis. In: *Proceedings Text, Speech and Dialog*, Pilsen, Czech Rep., LNAI 5729, (2009), 77-84.
30. Steinberger, J., Ježek, K.: Text Summarization: An Old Challenge and New Approaches. *Foundations of Computational Intelligence Vol 6*, Springer (2009), 127-149.
31. Steinberger, J., Ježek, K.: Evaluation Measures for Text Summarization. In: *Computing and Informatics*, 28(2), (2009), 1001-1025.
32. Steinberger, J., Ježek, K., Sloup, M.: Web Topic Summarization. In: *Comference on Electronical Publishing, ELPUB 08*, Toronto, Canada, (2008), 322-334.
33. Steinberger, J., Poesio, M., Kabadjov, M.,A., Ježek, K.: Two uses of anaphora resolution in summarization. In: *Information Processing and Management*, 43(6), (2007), 1663-1680.
34. Witte, R., Bergler, S.: Next-Generation Summarization: Contrastive, Focused and Update Summaries. In: *Conf. on Recent Advantages in Natural Language Processing, RANLP'07*, Borowetz, Bulgaria, (2007).
35. Yeh, J.,Y., at all: Text summarization using a trainable summarizer and latent semantic analysis. In: Information Processing and Management, 41(1), (2005), 75–95.

Annotation:

This paper introduces a taxonomy of summarization methods and an overview of their principles from single document over multi-document to advanced techniques as update and focused summarization. A special attention is devoted to application of recent information reduction methods, based on singular value decomposition. The last part is dedicated to evaluation of quality of a summary and various measures for its assessment are described.

Acknowledgement: This work was partly supported by grant no. 2C06009 Cot-Sewing.