Kontrastivní sumarizace textů

Michal Campr[1]

1 Úvod

Vlivem neustálého rozvoje internetu, jakožto zdroje informací, jsme zahlcování obrovským množstvím dat, ve kterých se buď nedokážeme, nebo z časových důvodů nemůžeme orientovat. V současné době je proto věnována velká pozornost metodám, které automaticky zredukují danou množinu dat při maximálním zachování informační hodnoty.

Mezi takové metody patří i automatická sumarizace, která je určena pro zpracování textových dat. Existuje nespočet různých druhů sumarizace. Některé pracují na bázi statistiky, jiné využívají grafové nebo algebraické metody a jejich výsledkem může být tzv. extrakt (souhrn vytvořený ze sekvencí slov původního textu) nebo abstrakt (souhrn z nově vytvořených vět). V tomto příspěvku se budu zabývat jednou konkrétní metodou vytvářející extrakty – automatickou kontrastivní sumarizací s využitím latentní sémantické analýzy.

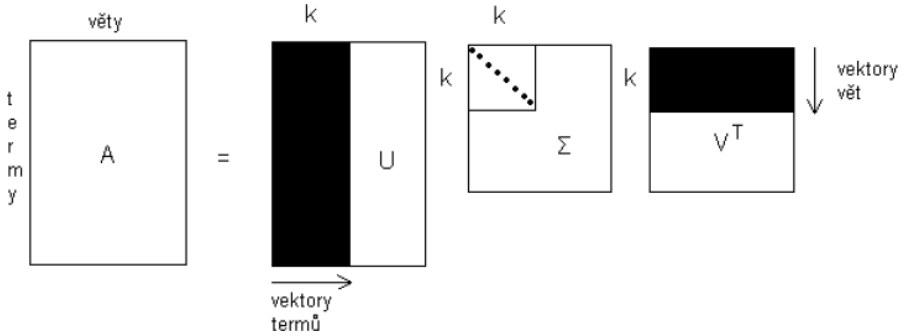
2 Latentní sémantická analýza

Latentní sémantická analýza (LSA) je algebraická metoda, která automaticky analyzuje vztahy mezi termy a větami pomocí tzv. singulární dekompozice matic (SVD – Singular Value Decomposition). Pro jednoduchost bude nyní LSA vysvětlena na vytvoření extraktu z jednoho dokumentu.

Základem je vytvoření matice A, kde sloupcové vektory A_i představují vektory frekvencí termů ve větě i vstupního dokumentu. Výsledkem je řídká matice $m \times n$, kde m je počet termů v dokumentu a n je počet vět (obr. 1). Singulární rozklad matice A je pak definován:

$$A = U\Sigma V^T, \tag{1}$$

kde U je $m \times n$ sloupcově ortonormální matice, jejíž sloupce se nazývají levé singulární vektory, Σ je $n \times n$ diagonální matice obsahující tzv. singulární čísla (seřazená sestupně) a V je $n \times n$ ortonormální matice obsahující pravé singulární vektory. Rozměry matic jsou pak redukovány na $k < n$ dimenzí.



Obrázek 1: Singulární dekompozice matice A

[1] Student doktorského studijního programu Inženýrská informatika, obor Informatika a výpočetní technika, specializace Sumarizace textů a její využití v multijazykovém prostředí webu, e-mail: mcampr@kiv.zcu.cz

Matice U pak mapuje termy do témat obsažených v dokumentu, Σ reprezentuje významnost témat a V^T mapuje věty do témat. Pro vytvoření extraktu tedy stačí vybrat z matice V^T požadovaný počet pravých singulárních vektorů a příslušné věty zařadit do souhrnu.

3 Kontrastivní sumarizace pomocí LSA

Výše popsanou techniku lze dále upravovat a přizpůsobovat daným potřebám, jako například vytváření souhrnu z více dokumentů najednou nebo porovnávání obsahu dokumentů. Kontrastivní sumarizace se zabývá právě tímto problémem a úzce souvisí s metodou aktualizační sumarizace popsanou v práci Steinberger a Ježek (2009).

Na počátku jsou dány dvě množiny dokumentů X a Y. Cílem je vytvoření dvou souhrnů, které reflektují rozdílnost obou množin dokumentů, tzn. co je v množině Y navíc oproti X a naopak. V duchu výše uvedené metody nejprve vytvoříme dvě matice A_X a A_Y odpovídající oběma množinám, s tím rozdílem, že je nutné brát v úvahu termy z obou množin dohromady. Pomocí SVD pak vytvoříme rozklad obou matic, čímž získáme matice U_X a U_Y, Σ_X a Σ_Y, V_X^T a V_Y^T. Následující postup provedeme pro vytvoření souhrnu, obsahujícího nejvýznamnější informace, které jsou v množině dokumentů Y navíc oproti X.

Pro každé téma t dané sloupcem matice U_Y vyhledáváme nejpodobnější (tj. redundantní) téma dané sloupcem U_X. Redundanci dvou vektorů udává kosinová podobnost:

$$red(t) = \max_{i=1}^{k} \frac{\sum_{j=1}^{m} U_X[j,i] * U_Y[j,t]}{\sqrt{\sum_{j=1}^{m} U_X[j,i]^2} * \sqrt{\sum_{j=1}^{m} U_Y[j,t]^2}} \quad (2)$$

Novost tématu je pak dána vztahem *1-red(t)* a v kombinaci s významností témat danou singulárními čísly $\sigma(t)$ v matici Σ pak získáme diagonální matici US (update score):

$$us(t) = \sigma(t) * (1 - red(t)) \quad (3)$$

Vynásobením $US \cdot V^T$ dostaneme matici F, která v sobě agreguje novost i důležitost nových témat. V této matici pak vyhledáváme věty, které mají nejdelší vektor a ty pak zařazujeme do souhrnu. Délka vektoru s_t pro větu t je dána:

$$s_r = \sqrt{\sum_{i=1}^{k} v_{ri}^2 * \sigma_i^2} \quad (4)$$

Po nalezení nejdelšího vektoru vynulujeme příslušný vektor v matici F a pokračujeme iteračně dál, dokud nemá souhrn požadovanou délku.

Stejným způsobem vytvoříme i druhý souhrn pro opačný směr porovnání.

3 Závěr

Kontrastivní sumarizace textů pomocí latentní sémantické analýzy je nová metoda, se kterou v současné době experimentuji a testuji její výkonnost. Dosavadní výsledky naznačují, že je pro daný problém dobře použitelná a zároveň dostatečně rychlá.

Literatura

Ježek, K., Steinberger, J., 2010. Sumarizace textů. *Proceedings of Annual Database Conference DATAKON*, Mikulov, Czech Rep., pp.3-23, ISBN 978-80-7368-424-2

Steinberger, J., Ježek, K. 2009. Update summarization based on latent semantic analysis. *In Proceedings of 12th International Conference*, TSD 2009, Pilsen, Czech Republic,