METHODS OF DOCUMENT SUMMARIZATION ON THE WEB

Michal Belica
Master Degree Programme (2), FIT BUT
E-mail: xbelic02@stud.fit.vutbr.cz

Supervised by: Vladimír Bartík
E-mail: bartik@fit.vutbr.cz

Abstract: The paper describes automatic summarization of web documents. Firstly previous work and current state are mentioned. Document preprocessing and conversion into representation suitable for summarization algorithms are briefly presented. The project is mainly focused on the automatic document summarization with the main attention paid to an advanced algorithm that uses latent semantic analysis.

Keywords: data mining, text summarization, data reduction, web-data extraction, Python, NLP, natural language processing, latent semantic analysis, LSA, singular value decomposition, SVD

1 ÚVOD

Pod pojmom automatická sumarizácia textu budem v tejto práci rozumieť proces vytvorenia súhrnu z textového dokumentu vo formáte HTML. Výsledný súhrn môže byť použitý napr. pre rýchle zoznámenie sa s niekoľkými dokumentami a pomôcť pri rozhodnutí o vhodnosti čítania ich celého obsahu. Dlhšie súhrny však môžu byť použité aj ako náhrada za pôvodný dokument.

Nasledujúca sekcia v krátkosti popisuje hlavné míľniky v metódach pre automatickú sumarizáciu a existujúce nástroje. Sekcia 3 sa podrobnejšie zaoberá mnou použitým prístupom k automatickej sumarizácii webových dokumentov. V záverečnej sekcii 4 sú zhrnuté poznatky z tejto práce.

2 SÚČASNÝ STAV

Medzi prvé pokusy o automatizované vytvorenie sumarizácie patrí Luhnova práca [5]. Ten využíval jednoduchú heuristiku, ktorá sa zakladala na tom, že najvhodnejšími vetami do súhrnu sú tie s najviac frekventovanými frázami. Ďalšou významnou prácou boli experimenty Edmundsona [1], ktorý prispel troma novými heuristikami a tým dosiahol zlepšenie kvality výsledných súhrnov. Tieto metódy, hoci boli jednoduché, vykazovali veľmi dobré výsledky. Časom sa ale začali objavovať sofistikovanejšie prístupy ako napríklad metódy založené na Naivnej Bayesovskej klasifikácii [4] alebo rôzne metódy z oblasti soft-computingu [3].

Jednou z moderných metód používaných v súčastnosti, nie len pre potreby sumarizácie, je metóda založená na LSA[1] [2, 3, 6]. Jej veľkou výhodou pri spracovaní textu je to, že implicitne analyzuje skryté vzťahy medzi slovami, slovnými spojeniami a vetami nezávisle na použitom prirodzenom jazyku. Tým sa dokáže efektívne vysporiadať s problémami ako sú napr. nejednoznačnosť výrazov a problém synonymie.

Myšlienka automatickej sumarizácie nie je nová, a preto existuje niekoľko funkčných implementácií. Aplikácia *Open Text Summarizer* (`http://libots.sourceforge.net/`) je implementácia dostupná na platforme UNIX, ktorá využíva jednu z pokročilejších sumarizačných techník.

[1] LSA - latentná sémantická analýza, angl. *Latent Semantic Analysis*

Modernými knižnicami sú *Musutelsa* (`http://www.musutelsa.jamstudio.eu/`) a *Almus* (`http://textmining.zcu.cz/?section=download`) využívajúce metódu LSA. Najkomplexnejším riešením je platforma *MEAD* (`http://www.summarization.com/mead/`), ktorá obsahuje aj nástroje určené pre vyhodnocovanie kvality sumarizácií. Každá z uvedených implementácií má však nejakú nevýhodu. Najčastejšou nevýhodou je nutnosť kompatibilného vstupného XML súboru, či podpora len niekoľkých jazykov.

3 POPIS SUMARIZAČNÉHO SYSTÉMU

Významnú časť sumarizácie tvorí predspracovanie, ktoré spočíva hlavne v prevedení dokumentu vo formáte HTML do modelu dokumentu vhodného pre ďalšie spracovanie. Automatickú extrakciu hlavného textu zo štruktúry HTML dokumentu zabezpečuje platforma *Readability*. Dokument je následne rozdelený na tokeny (slová, vety, ...), sú odstránené slová bez sémantického významu (tzv. stop-slová) a na zvyšné slová je aplikovaný stemming.

Sumarizačná metóda LSA vyžaduje aby bol dokument reprezentovaný vektorovým modelom. Pre jednotlivé zložky vektorového priestoru je použitá metrika uvedená v rovnici 1, kde $\mathrm{f}(t,d)$ vyjadruje počet termov t (slov) v dokumente d, $\mathrm{MaxFreq}(d)$ vyjadruje počet výskytov najfrekventovanejšieho termu v dokumente d a $0 \leq s < 1$ je parameter, ktorého úlohou je tlmiť príspevok druhého výrazu (zvyčajne $s = 0{,}5$). Metrika je podporená váhami pre slová obsiahnuté v HTML tagoch `<strong>`, `<a>`, ... tak aby sa zohľadnili autorom označené významné slová, prípadne slová, ktoré odbiehajú od témy. Tým sú ovplyvnené všetky vety, v ktorých sa tieto slová vyskytujú a vďaka metóde LSA aj všetky sémanticky podobné slová.

$$\mathrm{TF}(t,d) \quad = \quad s + (1-s)\frac{\mathrm{f}(t,d)}{\mathrm{MaxFreq}(d)} \tag{1}$$

Celý dokument, ktorý obsahuje m slov a n viet je možno vyjadriť ako maticu $A = [A_1, A_2, \ldots, A_n]$ rozmeru $m \times n$, kde riadky predstavujú slová a stĺpce vety dokumentu. Stĺpcový vektor A_i reprezentuje frekvencie slov vo vete i pôvodného dokumentu spočítané podľa rovnice 1. Maticová reprezentácia dokumentu je pomocou metódy SVD[2] rozložená na 3 matice znázornené na obrázku 1.

$$A = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 1 & \cdots & 1 \end{bmatrix} = \underset{U}{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ u_{21} & & & \\ & & & \\ \vdots & & \ddots & \\ & & & \\ & & & \\ u_{m1} & & & u_{mm} \end{bmatrix}} \underset{\Sigma}{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_r \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}} \underset{V^T}{\begin{bmatrix} v_{11} & \cdots & v_{1n} \\ v_{21} & & \\ \vdots & \ddots & \\ & & \\ & & \\ v_{n1} & & v_{nn} \end{bmatrix}}$$

Obrázek 1: Singulárny rozklad matíc s naznačeným redukovaným priestorom

Matica $U = [u_{ij}]$ rozmeru $m \times n$ je stĺpcovo ortogonálna a jej stĺpce sa nazývajú ľavé singulárne vektory. Diagonálna matica $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ rozmeru $n \times n$ obsahuje nezáporné singulárne čísla zoradené zostupne (diagonálne prvky). Ak r je rád matice A, potom platí vzťah 2. Nakoniec $V = [v_{ij}]$ je ortogonálna matica rozmeru $n \times n$, ktorej stĺpce sa nazývajú pravé singulárne vektory.

[2] SVD - singulárna dekompozícia (angl. *Singular Value Decomposition*)

Rozmer matíc je redukovaný na k dimenzií, kde $k < n$. Z toho vyplýva, že matice sú redukované nasledovne: U na $m \times k$, Σ na $k \times k$ a V^T na $k \times n$ (obrázok 1).

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0 \tag{2}$$

Na rozklad matice sa môžeme pozerať ako na rozklad pôvodného dokumentu do k lineárne nezávislých bázových vektorov reprezentujúcich hlavné témy textu. Výsledkom je, že podobná kombinácia slov sa bude vyskytovať pozdĺž rovnakého singulárneho vektoru. Potom matica A mapuje slová do jednotlivých viet a matice U a V mapujú slová resp. vety do k najvýznamnejších tém.

Sumarizačná metóda využíva rozklad matice V^T slov proti vetám. Matica popisuje mieru významnosti viet v témach dokumentu. Následne sa vyberie veta s najväčšou dĺžkou vektorovej reprezentácie $S = \Sigma^2 \times V^T$. Násobením Σ^2 sa zohľadní štatistická významnosť hlavných tém, ktorá je úmerná druhej mocnine príslušného singulárneho čísla. Do výsledného súhrnu sú následne zaradené vety, ktoré majú najvyššie hodnoty s.

4 ZÁVER

V práci bol predstavený systém pre automatickú sumarizáciu webových dokumentov. Pretože celý proces sumarizácie vrátane predspracovania je plne automatizovaný, tak nie je potrebná žiadna interakcia človeka. Na samotnú sumarizáciu je využitá moderná metóda, ktorá je nezávislá na jazyku dokumentu a dokáže sa vyrovnať s problémami, ktorými trpia iné sumarizačné metódy. Oproti bežne rozšíreným sumarizátorom sa metóda snaží ťažiť z meta-informácií o dokumente, ktoré ponúka formát HTML aby sa dosiahlo vyššej kvality výsledného súhrnu.

REFERENCIE

[1] Edmundson, H. P. New Methods in Automatic Extracting. J. ACM. Apríl 1969, roč. 16, č. 2. S. 264–285. Dostupné na: <http://doi.acm.org/10.1145/321510.321519>. ISSN 0004-5411.

[2] Gong, Y. a Liu, X. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2001. S. 19–25. SIGIR '01. Dostupné na: <http://doi.acm.org/10.1145/383952.383955>. ISBN 1-58113-331-6.

[3] Ježek, K. a Steinberger, J. Sumarizace textů. In DATAKON. 2010.

[4] Kupiec, J., Pedersen, J. a Chen, F. A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 1995. S. 68–73. SIGIR '95. Dostupné na: <http://doi.acm.org/10.1145/215206.215333>. ISBN 0-89791-714-6.

[5] Luhn, H. P. The automatic creation of literature abstracts. IBM Journal Res. Dev. Apríl 1958, roč. 2, č. 2. S. 159–165. Dostupné na: <http://dx.doi.org/10.1147/rd.22.0159>. ISSN 0018-8646.

[6] Steinberger, J. a Ježek, K. Using latent semantic analysis in text summarization and summary evaluation. In Proceedings ISIM '04. 2004. S. 93–100.