COIA-2008
BAKU, 2-4 June

The Second International Conference on Control and Optimization with Industrial Applications, COIA 2008, June 2-4, Baku

MULTIDOCUMENT SUMMARIZATION BY SENTENCE RANKING

Aliguliyev Ramiz M.

Institute of Information Technology of Azerbaijan National Academy of Sciences, a.ramiz@sciences.az

Occurrence of WWW at the end of the previous century has led to fast growth of information quantity, accessible to users. From all kinds of the information saved up on WWW, text data consist not less than 90 % of the information. As it is known, in order to find in such huge knowledge bases something valuable is possible only by means of specialized technologies. One of these technologies is text mining. Text mining technology, is developed on the base of the static and linguistic analysis, as well as, an artificial intellect, it is intended for content analysis and search in unstructured text data. With application of text mining technology users can receive valuable information - new knowledge. Basic functions of text mining can include the followings: summarization, clustering, classification, feature selection, question answering, thematic indexing, keyword searching, also creating taxonomies and thesauri.

Let given collection of documents $\mathbf{D} = \{D_1, D_1, ..., D_N\}$. Through $S = \{S_1, S_2, ..., S_M\}$ we shall denote a set of sentences in this collection. Sentences collection $S = \{S_1, S_2, ..., S_M\}$ we present as a graph where vertexes correspond to sentences, and the weight of edges corresponds a similarity measure between them. Let's admit, that each sentence is presented as a sequence of words, appearing in it,$S_j = \{t_1, t_2, ..., t_{m_j}\}$. $j = 1, 2, ..., M$. Then a similarity measure between pair $S_k \in S$ and $S_l \in S$ we shall determine as following formula:

$$sim(S_k, S_l) = \frac{|S_k \bigcap S_l|}{|S_k| + |S_l| - |S_k \bigcap S_l|}, \tag{1}$$

where $|S_k|$ denotes the number of words in the sentence S_k, $|S_k \bigcap S_l|$ is the number of identical words in both sentences S_k and S_l. From definition (1) follows that, if $|S_k \bigcap S_l| = |S_k| = |S_l|$, so $sim(S_k, S_l) = 1$. On the contrary, if $|S_k \bigcap S_l| = 0$, so $sim(S_k, S_l) = 0$.

The ranking algorithm WICER is given by formula:

$$PR_i(j) = \frac{1-d}{M} + d\left(1 + \frac{N_i}{N}\right) \sum_{p=1}^{N} W_p \sum_{q \in O_i^p} w_{pi} \frac{PR_p(q)}{Out(q)}. \tag{2}$$

Where,

(1) $PR_i(j)$ is the rank of sentence S_j of document D_i;
(2) $M$is the number of sentences in the document collection $D = \{D_1, D_1, ..., D_N\}$;
(3) N is the number of documents in the document collection $D = \{D_1, D_1, ..., D_N\}$;
(4) N_i is the number of documents that have an edge to document D_i;
(5) d is the damping factor, $d \in [0.8, 1]$;
(6) W_p is the weight of document D_p;
(7) O_i^p is the set of sentences in document D_p having links to sentence S_j of document D_i;
(8) w_{pi} is the of the edge from document D_p to D_i, $w_{pi} = \begin{cases} \alpha, & p \neq i \\ \beta, & p = i \end{cases}$ (parameters α and β are the inter document and intra-document edge weights, respectively).

Finally, as to selection of sentences to generate a summary, in each document sentences are sorted in reversed order of their score and the top ranked sentences are selected for in the extractive summary.