Usando Grafos e Algoritmo de Karp na Sumarização Automática de Documentos

Andréa Carneiro Linhares
Universidade Federal do Ceará
Praça Senador Figueira, rua Anahid Andrade, Centro, CEP 62.040-001 – Sobral/CE
andrea.linhares@ufc.br

Alejandro Molina Villegas, Juan-Manuel Torres Moreno
Université d'Avignon,
Laboratoire Informatique, BP 1228, Avignon Cedex 9 84911, France,
alemol@gmail.com, juan-manuel.torres@univ-avignon.fr

Peter Peinl
University of Applied Sciences Fulda
Faculty of Computer Science, Marquardstraße 35, Fulda 36039, Germany
peter.peinl@informatik.fh-fulda.de

RESUMO

A sumarização automática de documentos é um dos métodos de mineração de texto (*text mining*) que permite comprimir um documento com perda de informação, conservando sua informatividade. Os primeiros trabalhos sobre o assunto, problemática importante do Processamento de Linguagem Natural (PLN), datam do final dos anos 50. Eles possibilitaram a definição de uma metodologia de produção de sumários:

- Pré-processamento do texto;
- Identificação das frases em destaque no documento;
- Construção do sumário por concatenação das frases extraídas;

Os métodos desenvolvidos consideravam o resumo por extração como a identificação dos vértices mais importantes num grafo. Os algoritmos determinam a importância de um vértice (frase) baseando-se na análise local do mesmo e na informação obtida da análise recursiva do grafo. Em seguida, as frases são selecionadas segundo critérios de prestígio no grafo e finalmente reunidas para produzir os resumos.

Nesse contexto, nosso trabalho propõe-se modelar a sumarização automática de documentos por meio de grafos e utilizar algoritmos heurísticos ou exatos oriundos da Teoria dos Grafos no processo de classificação das frases (sentenças). A sumarização automática é definida como um problema de otimização, sem a restrição de tamanho (número de frases extraídas) do resumo fornecido. A primeira abordagem desenvolvida buscou não apenas classificar as frases, mas sim determinar automaticamente resumos de comprimento médio. Nesse intuito, modelamos o documento como um grafo em que vértices representam frases e arcos denotam as ligações possíveis entre as frases. Atribuímos um fator de peso aos arcos de acordo com a probabilidade de distribuição dos termos. Os algoritmos desenvolvidos foram baseados no algoritmo de Karp (pesquisa de um ciclo de peso médio mínimo). Eles utilizam uma matriz de pesos M, cujo elemento m_{ij} representa a soma da divergência de Jenssen-Shannon (JS) da frase "i" e o texto T com a divergência JS da frase "j" e T. Além disso, uma "melhoria" α é adicionada se a frase "i" ou "j" aparece no título do documento. Assim, uma solução é uma sequência de vértices (frases) que pertencem ao ciclo de peso médio mínimo. Algumas restrições foram introduzidas nas versões desenvolvidas: modelagem do problema usando um grafo orientado completo; exigência de que os circuitos sejam elementares, entre outras.

Os primeiros resultados mostraram-se promissores. Nós avaliamos os resumos produzidos por nossos métodos com o software ROUGE, que mede a similaridade entre um resumo candidato (produzido automaticamente) e resumos de referência (produzidos por humanos). Uma nova modelagem matemática para o problema baseada no Problema do Subconjunto Estável Máximo está sendo investigada e pretendemos comparar os resultados obtidos com o novo método.

PALAVRAS-CHAVE. Sumarização automática de documentos, Problema de Otimização, Algoritmo de Karp. Área principal: TAG.

ABSTRACT

Automatic summarization of documents is one of the text mining methods used for compressing a text with loss of information, while keeping its informativeness. The first works in this field, an issue important to the Natural Language Processing (NLP) date from the late 50s. They allowed the definition of a methodology to produce summaries:

- Pre-processing of the text;
- Identification of highlighted phrases in the document;
- Construction of the summary by concatenation of the phrases extracted;

The developed methods generate the summary by extracting the vertices of a graph identified as the most important ones. The algorithms determine the importance of a vertex (sentence) based on its local analysis and on the information obtained from the recursive analysis of the graph. Then, the sentences are selected according to a ranking criterion in the graph and finally assembled to produce the summaries.

In this context, the aim of our work is to propose a graph model for the automatic summarization of documents and apply heuristics or exact methods coming from graph theory in the sorting of the phrases (sentences). The automatic summarization is defined as an optimization problem without the size constraint (number of sentences extracted) of the summary provided. The first approach adopted sought not only to classify the sentences, but rather to automatically determine summaries of medium length. So, we model the document as a graph where vertices represent phrases and arcs denoting possible links between sentences. We assigned a weighting factor to the arcs according to the probability distribution of terms. The developed algorithms were based on the Karp algorithm (searching for a minimum average weighted length cycle). They use a weight matrix M, whose element m_{ij} represents the sum of the Jenssen-Shannon divergence (JS) of the phrase "i" and the text T with divergence JS of the phrase "j" and T. The weight is increased by α if the phrase "i" or "j" appears in the document title. Thus, a solution is a sequence of vertices (sentences) that belong to the cycle of minimum weight. Some constraints have been introduced into the developed versions, among others: modeling the problem as a complete direct graph; a requirement that the circuits are elementary.

The first results are promising. We evaluate the summaries produced by our methods with the software Rouge, which measures the similarity between a candidate summary (produced automatically) and summaries of reference (produced by humans). A new mathematical model of the problem based on the Maximum Stable Set Problem is being investigated and we intend to compare the results with the new method.

KEY WORDS: Automatic summarization of documents, Optimization Problems, Karp's algorithm. Main area: TAG.