Explorando Medidas de Redes Complexas para Sumarização Multidocumento

Ademar T. Akabane[1], Thiago A. S. Pardo[1], Lucia H. M. Rino[2]
[1]Instituto de Ciências Matemáticas e de Computação, USP
[2]Departamento de Computação, UFSCar

Abstract. *This paper explores the use of complex networks for multi-document summarization, which is the task of automatically producing a unique informative summary from a collection of texts on the same topic. We show that simple complex networks metrics may be useful for this task.*

Resumo. *Este artigo explora o uso de medidas de redes complexas para sumarização multidocumento, a qual consiste em produzir um único sumário a partir de uma coleção de textos que abordam o mesmo assunto. Mostramos que medidas simples de redes podem ser úteis para esta tarefa.*

1. Introdução

A cada ano que se passa, a quantidade de informações geradas e disponíveis em meio digital cresce em ritmo acelerado. Nesse cenário, a Sumarização Automática (SA) de textos está ganhando cada vez mais relevância. Na área de SA multidocumento, procura-se desenvolver técnicas que possibilitem extrair o conteúdo mais relevante de um ou mais textos-fonte de forma resumida, mantendo o significado original e com a mínima intervenção humana.

Recentemente, grafos e Redes Complexas (RC) (Newman, 2003) têm se mostrado ferramentas importantes para modelar textos e se abordar a questão da SA, pois são representações elegantes, poderosas e muito exploradas. As RC diferem dos grafos tradicionais principalmente pela grande quantidade de vértices, sua organização segundo princípios complexos (não aleatórios) e por apresentar características topográficas particulares que não se encontram em grafos mais simples.

Salton et. al. (1997), em seu trabalho pioneiro, introduziram o que é considerado hoje um dos principais trabalhos de sumarização com o uso de grafos. Mais recentemente, Mihalcea (2004) utiliza com sucesso o conceito de grafos para a sumarização com inspiração em algoritmos de recuperação de informação. Outros trabalhos importantes dessa área de RC aplicada a SA foram os de Antiqueira (2007) e de Leite (2010). O primeiro propôs 26 métodos de sumarização baseadas em 10 medidas numéricas e características de RC. Cada método foi utilizado individualmente, gerando diferentes modelos de SA. Já o segundo utilizou modelos de SA baseados na Teoria de Grafos, aprendizado de máquina e diversas estatísticas, buscando combinar e selecionar características diversas de SA de forma a explorar diferentes aspectos do texto. Esses trabalhos são para SA monodocumento.

O objetivo deste trabalho também é investigar métodos de sumarização com base em métricas e propriedades de grafos e RC, mas, diferentemente dos trabalhos anteriores, foca-se na SA multidocumento, abordando-se suas características próprias e desafios. A estrutura desse trabalho está organizada da seguinte maneira: na Seção 2, apresentam-se, de forma sucinta, os principais conceitos utilizados nesse trabalho; na Seção 3 é descrito o método de SA proposto.

2. Conceitos Básicos

As redes utilizadas no trabalho de Antiqueira (2007) foram construídas da seguinte maneira: cada texto é representado por uma rede, sendo que cada sentença representa um vértice e as arestas indicam a similaridade entre sentenças distintas, sendo esta similaridade calculada pela ocorrência de substantivos em comum.

As 10 medidas numéricas de RC usadas são: (i) grau, (ii) coeficiente de aglomeração, (iii) caminho mínimo, (iv) índice de localidade, (v) índice de concordância, (vi) grau hierárquico, (vii) *d*-anéis, (viii) *k*-núcleos, (ix) *k*-cortes e (x) comunidades. Em seguida serão descritos apenas três

medidas. Essas três medidas escolhidas são utilizadas nesse trabalho de iniciação científica. Foram escolhidas por sua simplicidade e bons resultados no trabalho acima.

1.*Grau:* O grau de um nó *i* (k_i, onde, $0 \leq k_i \leq N-1$ e *N* é o número total de nós da rede) representa a quantidade de arestas que ele possui, ou seja, é a quantidade de nós diferentes conectados a *i*. Essa medida indica o quão um nó é conectado com seus vizinhos. Portanto, quanto maior o valor do grau, maior será a sua informatividade. A equação é apresentada a seguir e a_{ij} representa a aresta do nó *i* conectado ao nó *j*:

$$k_i = \sum_{j=1}^{N} a_{ij} = \sum_{j=1}^{N} a_{ji}.$$

2.*Coeficiente de aglomeração:* Quando um vértice *i* está conectado a um vértice *j*, e o vértice *j* a um vértice *k*, essa medida verifica se há uma conexão entre os vértices *i* e *k*. O vértice que possui elevado coeficiente de aglomeração ($0 \leq C_i \leq 1$) está bem conectado a seus vizinhos, ou seja, há elevado compartilhamento de informações entre esses vértices. Para melhor entendimento dessa medida, considere que, para cada nó *i* da rede, existem k_i arestas que o associam a k_i outros nós. Se esses k_i nós formassem um clique, ou seja, se cada nó estivesse diretamente conectado a qualquer outro nó do conjunto, haveria k_i ($k_i - 1$)/2 arestas entre eles. Considerando-se que E_i é o número de arestas que realmente existem entre os k_i nós, apresenta-se abaixo a equação para o cômputo do coeficiente de aglomeração de um nó:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

3.*Caminho mínimo:* Um caminho mínimo que parte do nó *i* ao *j* é denotado por *dij,*ou seja, é a sequência mínima de arestas que leva um nó ao outro. Logo, se tomarmos a partir de um nó *i* da rede todos os caminhos mínimos possíveis, a todos os demais nós, tem-se a medida de distância média. Portanto, quanto menor esse valor de distância (sp_i), mais próximo, em média, o vértice *i* está dos demais, em outras palavras, mais informações relevantes ele possui. O cálculo de cada vértice é feito dentro da comunidade a que ele pertence (em que as comunidades representam redes desconexas). O valor do caminho mínimo ($1 \leq sp_i \leq N-1$, onde *N* é o número total de nós da rede) entre dois vértices é calculado pela equação mostrada abaixo:

$$sp_i = \frac{1}{N-1} \sum_{i \neq j} d_{ij} = \frac{1}{N-1} \sum_{i \neq j} d_{ji}$$

No trabalho de Leite (2010), por sua vez, foram propostos modelos de SA a partir da combinação ou do uso individual de três abordagens, são elas: grafos, aprendizado de máquina e estatísticas textuais. Ao todo, foram propostos 29 modelos de SA. Um dos focos principais do trabalho de Leite foi o aperfeiçoamento de uma ferramenta já existente de sumarização, originando o sumarizador SuPor-2.

Uma das conclusões relevantes no trabalho de Leite foi que a utilização de inúmeras características de SA não necessariamente melhora o desempenho do sumarizador. Por exemplo, a combinação do SuPor-2 com a RC obteve resultado inferior em relação ao SuPor-2 usado de forma isolada. Entretanto, a combinação de diferentes medidas obteve resultados melhores do que a aplicação individual das medidas proposta por Antiqueira.

3. Proposta de Método de SA Multidocumento

Neste trabalho, os textos a serem sumarizados são inicialmente representados em uma RC. A RC é construída da seguinte maneira: cada nó representa uma sentença e as arestas indicam relações/pesos entre pares de sentenças. Esses pesos são calculados a partir da tradicional medida do cosseno: quanto maior seu valor, maior será a similaridade entre as sentenças. As *stopwords* não são consideradas no cálculo da medida, pois, elas não contêm informatividade.

Além das três medidas de RC escolhidas e apresentadas anteriormente, outra medida foi acrescentada, que é a combinação das três medidas. A combinação é feita somando os valores de cada medida (antes de realizar a soma, é efetuada a normalização dos dados). As medidas utilizadas nesse trabalho associam um valor a cada nó da rede, dando, assim, embasamento à escolha das sentenças que devem compor o sumário.

A construção do sumário consiste em duas etapas: (a) seleção de sentenças e (b) remoção de sentenças redundantes (com relação às sentenças previamente selecionadas para o sumário). No último caso, utiliza-se o valor obtido no cálculo do cosseno, ou seja, duas sentenças são consideradas redundantes se têm cosseno maior que um valor V, onde V = (menor cosseno do grafo + maior cosseno do grafo) / 2. A escolha pelo valor médio de V deve-se principalmente ao fato de que existem variações no valor do cosseno de um conjunto de textos para outro, tornando inapropriado fixar um único valor. Essa duas etapas são controladas pela taxa de compressão especificada pelo usuário. Embora a taxa de compressão seja dada em número de palavras, somente sentenças completas são selecionadas.

Como ilustração, mostra-se, na Figura 4, um sumário produzido neste trabalho com a medida de caminho mínimo, com taxa de compressão de 70%.

> A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril.

Figura 4. Exemplo de sumário

Os sumários foram avaliados sobre o córpus CSTNews (Aleixo e Pardo, 2008), utilizando-se a medida ROUGE (Lin e Hovy, 2003). No geral, os melhores resultados foram obtidos pela medida de grau.

Agradecimentos

Ao PIBIC/CNPq e à FAPESP, pelo apoio a este trabalho.

Referências

Aleixo, P. e Pardo, T.A.S. (2008). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory.* Relatório Técnico 326, ICMC-USP. 12p.

Antiqueira, L. (2007). *Desenvolvimento de Técnicas Baseadas em Redes Complexas para Sumarização extrativa de Textos*. Dissertação de Mestrado. ICMC-USP.

Leite, D. S. (2010*). Um Estudo Comparativo de Modelos Baseados em Estatísticas Textuais, Grafos e Aprendizado de Máquina para a Sumarização Automática de Textos em Português*. Dissertação de Mestrado. DC-UFSCar.

Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram cooccurrence Statistics. In the *Proceedings of the Language Technology Conference*.

Mani, I. and Bloedorn, E. (1997). Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, Vol. 1, pp. 35-67.

Mihalcea, R. (2004). *Graph-based ranking algorithms for sentence extraction, applied to text summarization*. In the *Proceedings of the Annual Meeting of the Associaton for Computational Linguistics.*

Newman, M.E.J (2003). *The Structure and Function of Complex Networks*, SIAM Review 45, 167-256, cond-mat/0303516.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

Salton, G.; Singhal A.; Mitra, M; Buckley C. (1997). Automatic Text Structuring And Summarization. *Information Processing & Management*, Vol. 33, No, 2, pp. 193-207.