Multi-document Summarization with Graph Metrics

Rafael Ribaldo, Ademar Takeo Akabane, Thiago Alexandre Salgueiro Pardo

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

{ribaldo, takeusp}@grad.icmc.usp.br, taspardo@icmc.usp.br

Abstract. In this paper we introduce two systems - RSumm and CNSumm, which are multi-document summarizers based on the adaptation of the single-document relationship map and complex network methods, which represent texts as graphs and select sentences to compose the summary by using different graph traversing strategies and complex networks measures.

Keywords: summarization, relationship maps, complex networks, graphs

1 Introduction

Nowadays, with the huge and growing amount of information available in the Web and the sparse time to read and grasp it, manual analysis of the conveyed documents becomes almost impossible. For this reason, multi-document summarization has become an important area.

In this paper, we introduce two multi-document summarization (MDS) systems, called RSumm and CNSumm, which automatically produce a unique summary from a group of texts on the same topic for Brazilian Portuguese. RSumm adapts the traditional single-document relationship map methods [3], while CNSumm uses some complex network measures from the work of Antiqueira [1]. Complex Networks (CN) are more complex types of graphs. In contrast to simple graphs, CN present structures that tend to depart from random organization [2]. Both summarizers model texts as an undirected graph and different graph traversing strategies and measures may be used to select sentences to compose the summary.

2 Relationship Maps and Complex Networks Adaptation for MDS

Salton et al. [3] propose the representation of a text as a graph/relationship map in the following way: each paragraph becomes a node and the weighted edges are established among paragraphs with some lexical similarity (according to some lexical similarity measure, as word overlap and cosine measure). Only the best weighted edges are kept in the graph. Then, three possible methods for traversing the graph and selecting the paragraphs are proposed: bushy path, depth-first path and segmented bushy path. In the bushy path, the density, or *bushiness*, of a node is defined as the number of connections it has to other nodes. So, a highly linked node has a large overlapping vocabulary with several sentences, representing an important topic in the text. For this reason, it is a candidate for inclusion in the summary. Selection of highly connected nodes is done until compression rate is satisfied. In this way, the coverage of the main topics of the text is very likely to be good. However, the summary may be non-coherent, since relationships between every two nodes are not

properly tackled. To overcome that, instead of simply selecting the most connected nodes, the depth-first path starts with some important node (usually the most connected one) and continues the selection with the nodes (i) that are connected to the previous selected one and (ii) that come after it in the text, also considering selecting the most connected one among these, trying to avoid sudden topic changes. This procedure is followed until the summary is built. Its advantage is that more legible summaries may be built due to choosing sequential sentences. However, topic coverage may be damaged. The segmented bushy path aims at overcoming the bottlenecks introduced by the other two methods: it tackles the topic representation problem by first segmenting the graph in portions that may correspond to the topics of the text. Then, it reproduces the bushy path method in each subgraph. It is guaranteed that at least one paragraph of each topic will be included in the summary.

In this work, to build the graph (for both systems), the whole group of texts to be summarized is represented in a unique graph: all the sentences (instead of paragraphs) are represented in nodes and weighted edges are established among sentences with some lexical similarity (according to the cosine measure, after pre-processing the sentences – stemming and stopwords removal).

In relation to the RSumm tool, only the best weighed edges are kept in the graph and the bushy or the depth-first paths may be followed to select sentences to compose the summary.

In the case of CNSumm, degree, clustering coefficient and shortest path measures are used, following the results obtained by Antiqueira [1]. The well-known degree measure indicates how many connections one node has to the others. It is assumed that the higher the degree of a node, the more important the corresponding sentence is. The clustering coefficient measure signals how nodes tend to cluster together. It may indicate central topics to the summary. Also well-known, the shortest path measure indicates the length of the shortest path between 2 nodes. We use the average of the lengths of the shortest paths from a node to every other node in the network as an indication of its importance: the nearer a node is (in average) to the other nodes, the better the sentence is to compose the summary. Using one of the above CN measures, sentences are ranked from the best to the worst scored ones (according to the measure used). Then, starting from the first sentence in the rank, as many sentences as possible (according to some specified compression rate) are selected to compose the summary.

It is also necessary to treat redundancy in MDS for both systems, since it is a usual multi-document phenomenon. Therefore, a selected sentence is only included in the summary if it does not present a high lexical similarity value to any of the sentences previously selected.

It is interesting to notice that the summaries are built by simply juxtaposing the selected sentences. Future work must consider dealing with post-processing operations, as sentence ordering and fusion.

3 The Summarizer

Figure 1 shows two short texts and an automatically produced summary. For this example, the two paths (RSumm) and the Degree (CNSumm) produced the same summary. One may see that the summary is good.

Both systems are currently customized for Portuguese, since the stemmer and stoplist are for this language. However, it is important to notice that the proposed methods are language independent.

The demonstration of both systems will be done with a laptop, where any user may retrieve some texts from the web and test the system.

Text 1. *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. A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico. Por isso, Jade participará do evento em Buenos Aires, na Argentina, única cidade da América do Sul a receber o símbolo dos Jogos. O revezamento terminará em 8 de agosto, primeiro dia das Olimpíadas de Pequim.*

Text 2. *Um dos destaques desta temporada do esporte brasileiro, a ginasta Jade Barbosa foi escolhida, na noite desta terça-feira, para ser a representante do Brasil no revezamento da tocha dos Jogos Olímpicos de Pequim. Em votação pela internet, a ginasta recebeu mais de 100 mil votos e superou o nadador Thiago Pereira, que ganhou seis ouros nos Jogos Pan-Americanos. O Brasil não faz parte do trajeto da tocha olímpica. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril. Aos 16 anos, Jade conquistou três medalhas no Pan: ouro na disputa dos saltos, prata na apresentação por equipes e bronze no solo. Ao todo, a chama olímpica percorrerá 20 países antes de chegar a Pequim para a abertura da competição, no dia 8 de agosto.*

Summary. *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.*

Figure 1. Texts and their automatic summary

Acknowledgments

The authors are grateful to FAPESP and CNPq for supporting this work.

References

1. Antiqueira, L. (2007): *Desenvolvimento de Técnicas Baseadas em Redes Complexas para a Sumarização Extrativa de Textos*. MSc Dissertation. Instituto de Ciências Matemática e de Computação, Universidade de São Paulo. March, São Carlos/SP, Brazil, pp. 124
2. Barabási, A.L. (2003): *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume, New York
3. Salton, G., Singhal, A., Mitra, M., Buckley, C. (1997): *Automatic Text Structuring And Summarization*. Information Processing & Management, Vol. 33, N. 2, pp. 193-207