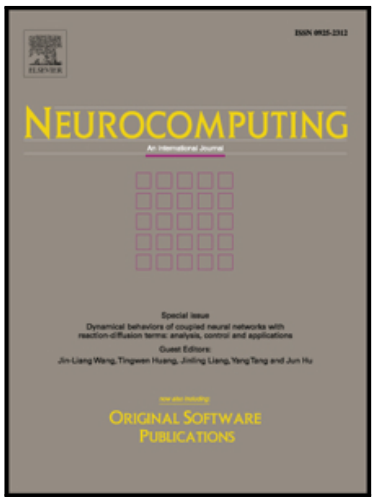

Multi-factors based Sentence Ordering for Cross-document Fusion from Multimodal Content

Lin Yue , Zhenkun Shi , Jiayu Han , Sen Wang , Weitong Chen , Wanli Zuo

PII: S0925-2312(17)30445-9
DOI: 10.1016/j.neucom.2016.12.084
Reference: NEUCOM 18195

To appear in: *Neurocomputing*

Received date: 29 April 2016
Revised date: 10 September 2016
Accepted date: 7 December 2016

Please cite this article as: Lin Yue , Zhenkun Shi , Jiayu Han , Sen Wang , Weitong Chen , Wanli Zuo , Multi-factors based Sentence Ordering for Cross-document Fusion from Multimodal Content, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2016.12.084

Multi-factors based Sentence Ordering for Cross-document Fusion from Multimodal Content

Lin Yue[1, 2, 3, 4], Zhenkun Shi[3, 6], Jiayu Han[3, 6], Sen Wang[5], Weitong Chen[4], Wanli Zuo[3, 6, *]

[1]School of Computer Science and Information Technology, Northeast Normal University, Changchun 130024, China
[2]School of Environment, Northeast Normal University, Changchun 130024, China
[3]Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changchun, Jilin, 130012, P. R. China
[4]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia
[5]School of Information and Communication Technology, Griffith University, Australia
[6]College of Computer Science and Technology, Jilin University, Changchun, Jilin, 130012, P. R. China
*Corresponding author

Abstract

Organizing a coherent structure of the sentences extracted from multiple documents, guarantees the fluency and readability of the fused document. In this paper, sentence ordering problem is treated as a combinatorial optimization problem and solved with continuous Hopfield neural network (CHNN). We unify the existing factors by considering the most frequent orders temporal information, and topical relevance between local themes during overall ordering process. Specifically, ordering algorithm traverses all the local themes and locates a shortest path as the final sentence ordering. We show the results with data from Document Understanding Conferences (DUC) 2002-2005, and demonstrate the effectiveness of the developed approach compared with Random Ordering (RO), Chronological Ordering (CO), Majority Ordering (MO), and Precedence Relation Ordering (PRO).

Keywords: sentence ordering; document fusion; continuous Hopfield neural network(CHNN)

1. Introduction

A system of interlinked hypertext documents accessed via the Internet constitutes the World Wide Web. With a web browser, one can view web pages that may contain videos, images, text, and other form of multimedia; and it allows us to navigate different content via hyperlinks. As the online-corpus is gigantic in its volume, Web search engines often return more results than actual needs. Navigation through all returned Web documents to obtain targeted information is infeasible and tedious burden, thus the automatic document summarization has been proposed for salient information retrieval [1, 2, 3, 4] and high-efficiency knowledge acquisition, which aims to produce a shortest description containing the most important information within all documents. Document fusion as a relevant research, aims to produce a shortest description with all information contained in the document sets, but without repetition [5, 6]. The significant difference is that, the former is like the intersection of document set, and the latter is the union of document set.

Both automatic document summarization and document fusion tackles the information overload problem in heterogeneous Web resources by providing a condensed and comprehensive version of a set of documents. Several key sub-tasks are involved in the research areas, such as redundancy removal [7], topic detection [8, 9], sentence retraction [10], objects merging in document set [11, 12], ordering sentence from different sources for keeping the logical and grammatical structure correct. Among all these extra tasks, sentence ordering is mandatory to compose sentences extracted from multiple documents into a coherent structure, which guarantees the fluency and readability of the results. The correct order of these sentences is helpful for understanding of the input articles. Moreover, the problem of information ordering is not limited to the areas mentioned hereinbefore, and concerns natural language generation (NLG) [13] applications such as in discourse planning and sentence aggregation [14, 15], which are important components of NLG. Besides, a brief, well-organized, fluent answer to a need for information at the specified level of granularity is also applicable in real-world question answering system, which is a classical application in social search. While it is trivial to order sentence from one single document, usually the extracted sentences are arranged as same order as in the original documents. The problem of sentence ordering for summarization or document fusion has received relatively little attention. The case we focus on is how to arrange the sentences extracted from different documents under a particular topic. It is a very important, but also potentially a very challenge task. As the sentences from source articles for ordering are written by different groups, from different viewpoints, or have different writing style, etc. There is not only the problem of subjective factors to deal with, but also the problem such as detecting rhetorical relations existing between sentences. It is difficult and unsolved in sentence ordering task. However, inferring a coherent ordering of extracted sentences with rhetorical structure analysis is not yet achievable.

Existing work on sentence ordering can be classified into two types: temporal information processing[16-18], and natural order learning in original corpora [19-23]. Among these methods, temporal information processing is the bottleneck technology which