Chapter 12

Combining Semantics and Social Knowledge for News Article Summarization

Elena Baralis
Politecnico di Torino, Italy

Saima Jabeen
Politecnico di Torino, Italy

Luca Cagliero
Politecnico di Torino, Italy

Alessandro Fiori
Institute for Cancer Research at Candiolo (IRCC), Italy

Sajid Shah
Politecnico di Torino, Italy

ABSTRACT

With the diffusion of online newspapers and social media, users are becoming capable of retrieving dozens of news articles covering the same topic in a short time. News article summarization is the task of automatically selecting a worthwhile subset of news' sentences that users could easily explore. Promising research directions in this field are the use of semantics-based models (e.g., ontologies and taxonomies) to identify key document topics and the integration of social data analysis to also consider the current user's interests during summary generation. The chapter overviews the most recent research advances in document summarization and presents a novel strategy to combine ontology-based and social knowledge for addressing the problem of generic (not query-based) multi-document summarization of news articles. To identify the most salient news articles' sentences, an ontology-based text analysis is performed during the summarization process. Furthermore, the social content acquired from real Twitter messages is separately analyzed to also consider the current interests of social network users for sentence evaluation. The combination of ontological and social knowledge allows the generation of accurate and easy-to-read news summaries. Moreover, the proposed summarizer performs better than the evaluated competitors on real news articles and Twitter messages.

DOI: 10.4018/978-1-4666-6086-1.ch012

INTRODUCTION

Since large document collections (e.g., news articles, scientific papers, blogs) are nowadays easily accessible through Web search engines, digital libraries, and online communities, Web users are commonly interested in exploring easy-to-read text summaries rather than perusing tens of potentially large documents. Multi-document summarization focuses on automatically generating concise summaries of large document collections. Text summarizers can be classified as sentence- or keyword-based. Specifically, sentence-based approaches entail partitioning document(s) into sentences and selecting the most informative ones to include in the summary (Carenini et al. 2007; Goldstein et al. 2000; Wang & Li 2010; Wang et al. 2011), whereas keyword-based approaches focus on detecting salient keywords to summarize documents using either co-occurrence measures (Lin & Hovy, 2003) or Latent Semantic Analysis (Dredze et al., 2008). Summarizers can be further classified as query-based or generic. While query-based summaries are targeted at a specific user query, the generic summarization task entails producing a general-purpose summary that consists of a selection of most informative document sentences or keywords. This chapter addresses the sentence-based generic multi-document summarization problem, which can be formulated as follows: given a collection of news articles ranging over the same topic, the goal is to extract a concise yet informative summary, which consists of most salient document sentences.

To effectively address document summarization, different data mining and information retrieval techniques have been adopted. For example, clustering techniques (e.g., Wang & Li, 2010; Wang et al., 2011) have been applied to first group document sentences into homogenous clusters and then pick out the most representative one within each cluster, whereas graph-based approaches (e.g., Radev, 2004; Thakkar et al., 2010) generate graphs that represent the underlying correlations between keywords or sentences. These models are then indexed by means of established graph ranking algorithms (e.g., Brin & Page, 1998) to identify the most salient document content. Unfortunately, general-purpose summarization strategies hardly differentiate between relevant concepts and not within a specific knowledge domain. Hence, the generated summaries may not meet reader's expectations and interests. To address this issue, two promising research directions have recently been investigated:

1. The exploitation of advanced semantics-based models (e.g., ontologies, taxonomies) to drive the summarization process (Conroy et al., 2004; Hennig et al., 2008; Wu & Liu, 2003); and
2. The integration of social data analysis steps to identify the current user interest's (e.g., Yang et al., 2011; Zhu et al., 2009).

Semantics-based approaches evaluate the document content according to established semantics-based models, such as ontologies or controlled vocabularies. Integrating ontologies into document summarizers allows us to automatically and effectively differentiate between terms having different meanings in different contexts as well as map term occurrences to their actual (non-ambiguous) concepts. The parallel analysis of the User-Generated Content (UGC) acquired from social networks and online communities can significantly improve summarizer performance (Conrad et al., 2009; Saravanan & Ravindran, 2010; Sharifi et al., 2010a; Yang et al., 2011; Zhu et al., 2009). For example, highlighting the current social trends (Mathioudakis & Koudas, 2010), the subjects that are currently matter on contention on the Web (Gong et al., 2010; Miao & Li, 2010), or the context in which Web documents were published (Yin et al., 2009) can be useful for generating appealing text summaries.