Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 43 (2007) 327–343

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Cross-document event clustering using knowledge mining from co-reference chains

June-Jei Kuo, Hsin-Hsi Chen *

Department of Computer Science and Information Engineering, National Taiwan University,
No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan

Received 16 May 2006; accepted 25 July 2006
Available online 11 October 2006

Abstract

Unifying terminology usages which captures more term semantics is useful for event clustering. This paper proposes a metric of normalized chain edit distance to mine, incrementally, controlled vocabulary from cross-document co-reference chains. Controlled vocabulary is employed to unify terms among different co-reference chains. A novel threshold model that incorporates both time decay function and spanning window uses the controlled vocabulary for event clustering on streaming news. Under correct co-reference chains, the proposed system has a 15.97% performance increase compared to the baseline system, and a 5.93% performance increase compared to the system without introducing controlled vocabulary. Furthermore, a Chinese co-reference resolution system with a chain filtering mechanism is used to experiment on the robustness of the proposed event clustering system. The clustering system using noisy co-reference chains still achieves a 10.55% performance increase compared to the baseline system. The above shows that our approach is promising.

Keywords: Controlled vocabulary; Co-reference chains; Event clustering; Multi-document summarization

1. Introduction

News through the Internet is an important information source, is reported anytime and anywhere, and is disseminated across geographic barriers. Detecting the start of new events and tracking their progress (Allan, Carbonell, & Yamron, 2002; Chen & Ku, 2002; Chieu & Lee, 2004) are useful for decision-making in today's fast-changing network era. The research issues behind event clustering include: how many features are used to determine event clusters, which cue patterns are employed to relate news stories in the same event, how clustering strategies affect clustering performance using retrospective data or on-line data, how the time factor affects clustering performance, and how cross-document co-references are resolved.

Several studies, for example, text classification (Kolcz, Prabakarmurthi, & Kalita, 2001) and web-page classification (Shen, Chen, Yang, Zhang, & Lu, 2004), suggest that even simple summaries are quite effective in

* Corresponding author. Tel.: +886 2 33664888x311; fax: +886 2 23628167.
E-mail addresses: jjkuo@nlg.csie.ntu.edu.tw (J.-J. Kuo), hhchen@csie.ntu.edu.tw (H.-H. Chen).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.07.016

carrying over relevant information about a document. They showed that if a full-text classification method is directly applied to those documents, it incurs much bias for the classification algorithm, potentially losing focus on the main topic and important content. Moreover, for deeper document understanding, the co-reference chains (Cardie & Wagstaff, 1999) of documents capture information on co-referring expressions, i.e., all mentions of a given entity. Since the co-reference provides important clues to find text fragments containing salient information, various practical tasks can be done more reliably, i.e., text summarization (Azzam, Humphreys, & Gaizauskas, 1999; Chen, Kuo, Huang, Lin, & Wung, 2003), question answering (Lin, Chen, Liu, Tsai, & Wung, 2001; Morton, 1999), event clustering (Kuo & Chen, 2004), etc. In contrast, while producing summaries from multiple documents, cross-document co-reference analyses (Bagga & Baldwin, 1998; Gooi & Allan, 2004) continue their consideration if there are the same mentions of a name in different documents.

This paper shows that using summarization as pre-processing in event clustering is a viable and effective technique. Furthermore, we integrate co-reference chains from more than one document by unifying cross-document co-references of nominal elements. Instead of using the traditional clustering approaches, we propose a novel threshold model that incorporates time decay function and spanning window to deal with on-line streaming news. The rest of the paper is organized as follows. Section 2 reviews the previous work and shows our architecture. Section 3 describes a document summarization algorithm using co-reference chains. Section 4 tackles the issues surrounding mining controlled vocabulary. A normalized chain edit distance and two algorithms are proposed to incrementally mine controlled vocabulary from cross-document co-reference chains. Section 5 proposes an algorithm for on-line event clustering using dynamic threshold model. Section 6 specifies the data set and the experimental results, using the metric adopted by Topic Detection and Tracking (Fiscus & Doddington, 2002). A Chinese co-reference resolution system is introduced in Section 7, a chain filtering algorithm is proposed to improve the quality of auto-tagged co-reference chains and the related experimental results are shown. Finally, Section 8 is a conclusion.

2. Basic architecture

Kuo and Chen (2004) employed co-reference chains to cluster streaming news into event clusters. They think the co-reference chains and event words are complementary in some ways, hence they also introduced the event words as defined by Fukumoto and Suzuki (2000). Kuo and Chen's (2004) experimental results showed that both factors are useful. Furthermore, they present two approaches to combine the two factors for event clustering, which are called summation model and two-level model. The summation model simply adds the scores for both co-reference chains and event words. On the contrary, a two-level model is designed in such a way that the co-reference chains or the event words are used separately rather than simultaneously. However, the best performance was by the summation model and improved only 2%, in terms of detection cost, compared to the baseline system. One of the reasons is that the nominal elements used in cross-document co-reference chains may be different. The goal of this paper is to mine, incrementally, controlled vocabulary from co-reference chains of different documents for event clustering on streaming news.

Fig. 1 shows the architecture of event clustering. We receive documents from multiple Internet sources, such as newspaper sites, and then send them for document pre-processing. The pre-processing module deals with the sentence extraction and language idiosyncracy, e.g., Chinese segmentation and co-reference resolution. Document Summarization module analyzes each document and employs the co-reference chains and the related feature words, such as event words or high TF-IDF words, to produce the respective summaries. The controlled vocabulary mining module integrates the co-reference chains to generate controlled vocabulary automatically. Finally, the event clustering module uses weights of word features, and a similarity function to cluster the documents.

3. Document summarization using co-reference chains

Kuo and Chen (2004) only used the event words as features for clustering. The basic hypothesis is that an event word associated with a news article appears across in a number of paragraphs, but a topic word does not. Moreover, the domain dependency among words is a key clue to distinguish a topic and an event. This can be captured by *dispersion value* and *deviation value* (Fukumoto & Suzuki, 2000). The former tells if a word