Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 365–378
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Use of place information for improved event tracking

Yun Jin [a], Sung Hyon Myaeng [b,*], Yuchul Jung [b]

[a] Chungnam National University, Daejeon 305-764, South Korea
[b] Information and Communications University, Daejeon 305-732, South Korea

Received 27 May 2006; accepted 25 July 2006

Abstract

The main purpose of topic detection and tracking (TDT) is to detect, group, and organize newspaper articles reporting on the same event. Since an event is a reported occurrence at a specific time and place and the unavoidable consequences, TDT can benefit from an explicit use of time and place information. In this work, we focused on place information, using time information as in the previous research. News articles were analyzed for their characteristics of place information, and a new topic tracking method was proposed to incorporate the analysis results on place information. Experiments show that appropriate use of place information extracted automatically from news articles indeed helps event tracking that identify news articles reporting on the same events.

Keywords: Place information; Event tracking; TDT; News article

1. Introduction

It is time consuming and laborious for individuals to detect a new event and track news articles reporting on the particular event from a variety of resources (Allan, 2002; Nallapati, Feng, Peng, & Allan, 2004). Topic detection and tracking (TDT) (Allan, Carbonell, Doddington, Yamron, & Yang, 1998) attempts to automate the process and provide a topic that consists of a seminal event or activity, along with all directly related events or activities. Here an event is a reported occurrence at a specific time and place and the unavoidable consequences, whereas an activity is a connected set of actions that have a common focus or purpose (Yang, Pierce, & Carbonell, 1998). Unlike traditional information retrieval or filtering, TDT focuses not just on topicality but on events or activities which often occur at a specific time and specific place (Allan, 2000; Nallapati et al., 2004).

Time information has been used in TDT. Based on the observation that articles reporting on a particular event tend to appear within a time window of two weeks, publication dates of news articles were used in topicality-based document clustering (Mani, Schiffman, & Zhang, 2003). More recent approaches attempted to

* Corresponding author.
E-mail addresses: myaeng@icu.ac.kr, wkim@cnu.ac.kr (S.H. Myaeng).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.07.007

automatically extract time information embedded in news article text and use it in determining whether the event described in an article refers to the event being tracked (Kim et al., 2004; Mani et al., 2000).

Place information seems to be as important as time in identifying an event, but it has neither been used nor tested extensively for usefulness in TDT. Possible reasons are:

- Time information such as publication dates is easily available from news articles, but place information needs to be extracted from news article text.
- Time information is one-dimensional and thus relatively easy to represent and compare, but place information is at least two-dimensional and more complex to determine its relationship with others. If place information is represented in a hierarchical way, for example, we should consider not only a parent–child relationship (e.g., Korea–Seoul) but also a sibling relationship (Seoul–Incheon).
- Compared to time information processing, place names seem to be more difficult to process because of the variety of expressions and ambiguities (e.g., Washington can be a person name or a state name, and there are several places with the name Rome in world geography).

Since our current work focuses on the use of place information for TDT, not on the techniques for extracting place information, we use existing techniques and resources for place name recognition. For instance, we made use of some portal sites, such as Naver,[1] Paran[2] and ESRI.[3] The sites provide services related to place name information, such as hierarchical information and/or location on a map for a given place name. This service is useful for determining the spatial relationship between two place names.

For effective use of place information, it is important to understand the characteristics in an event description. For instance, the place name as well as the time information tends to appear in the first sentence of a news article describing an event. By giving different weights for place names occurring in different parts of a new article, we can increase the chance of associating a place name with an event.

In sum, the contributions made in the research are as follows:

- We analyzed news articles to find important characteristics of how place information is used.
- We developed a novel event tracking method that incorporates the characteristics of place information.
- We demonstrated the value of place information in event tracking.

2. Related work

Research on distinguishing place and organization names as part of named entity recognition was reported in Message Understanding Conference (MUC)-6 (Sundheim et al., 1995). With the goal of better utilizing place information, a work on classifying such place information into categories like "city", "country", "region", and "water" was reported (Chinchor, 1998; Chung, Lim, Hwang, & Jang, 2004; Lee & Lee, 2004). Categorization of place names helps determining whether two names refer to the same location. For instance, tagging a place name "Geneva, New York" with [location] and [location], respectively, is less useful than tagging them with [city] and [state], since the latter can tell the relationship between "Geneva, New York" and "New York" with the aid of a place entity hierarchy. In other words, for example, "the accident in the state of New York" and "the accident in Geneva" may be considered the same event with a detailed place name tagging.

A few papers addressed the issue of using place information to improve effectiveness of TDT. Smith (2002) used place information for tracking similar historical events in unstructured history documents, not news paper articles. Documents containing the same dates and places, which were identified by named entity recognition, were considered as describing the same historical event. Juha, Helena, and Marko (2004) extracted person names, places, temporal expressions, and general terms to form four vectors, each representing one of the four types. The similarity between two documents was computed based on four vectors that were treated

[1] http://naver.com/.
[2] http://cube.paran.com/.
[3] http://maps.esri.com/scripts/esrimap.dll?name=gaz_avd&cmd=start/.