# International News Connection: A Real-time Online News Filtering and Classification System

**Zhiping Zheng**
School of Information
University of Michigan
zzheng@umich.edu

## ABSTRACT

Using formal information retrieval methods, International News Connection ([4]) provides a centralized location on the Web that allows users to access constantly updated international news, through dynamic links to news stories from 14 different news sources. The links are updated every 15-20 minutes. The news stories are classified into seven regional categories: Africa, Asia (excluding China), China, Europe (excluding Russia), Middle East, Russia, and South America. The system also contains a clustering function allowing users to retrieve news stories similar to a particular one.

**Keywords:** Web news, online news, dynamic information retrieval, automated classification, news filtering, news clustering, news updating behavior

## 1 INTRODUCTION

The classification and synthesized retrieval of the large amount of news on the Web has been a task that has attracted much research effort (e.g., [1], [2], [3]). The largest drawback of these news classification and retrieval systems lies at the fact that they are all based on static corpora of published news articles.

International News Connection is an attempt to build a dynamic news classification and retrieval system that allows users to access constantly updated news from different sources. It automatically creates constantly updated links to international news stories from 14 news sources, including, Agence France Presse, Associated Press, BBC, China Daily, CNN, International Herald Tribune, The Los Angeles Times, Nando Times, People's Daily, Reuters, The Sydney Morning Herald, USA Today, The Washington Post, and Xinhua News Agency. According to a self-adjusting updating schedule, a news fetcher will retrieve pre-selected index pages from news sources. The filtering function will extract relevant and new URLs from the index pages. Using the extracted URLs, the system will retrieve news stories, and make classification. For the purpose of clustering, a keyword vector is created and locally stored for each retrieved news story.

Figure 1 presents the processing diagram of International News Connection.

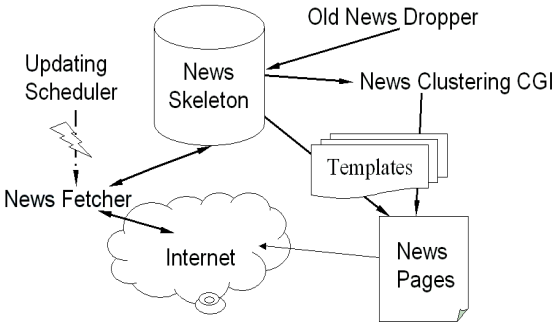


Figure 1. Processing diagram

## 2 CLASSIFICATION

For each of the seven international news categories, a feature vector is created from: 1) meronym from WordNet; 2) corpus training from content of selected Web pages; 3) manual selection, for example, word "Kursk" for Russia. In a feature vector, different words are given different weights. The weights are determined through experiments so that optimal classification outcome is achieved. To classify a news story, the cosine similarity method is used to compare the keyword vector of the news story to the feature vectors. Different threshold values are used for different categories.

## 3 FILTERING

The system conducts two kinds of filtering tasks. 1) From the index page of each news source, the system filters out non-news pages, obsolete news stories and special items, for example, News Brief; 2) From each retrieved news HTML file, it filters out content that is not a part of the news story.

The first filtering task is done mainly by pattern matching in URL. For example, each URL of news story links in The Sydney Morning Herald web site has following matching pattern: http://www.smh.com.au/news/YYMM/DD/*.html with YY, MM, and DD represent two-digit numbers of year, month, and day, and "*" represents category and filename in the URL. This filtering is also accomplished by other means, such as special title recognition.

For the second filtering task, different algorithms are adopted for different news sources, as they have different HTML page structures. All news Web pages contain contents that are not about the story itself, such as, HTML tags, JavaScript code, related story links, images, copyright information, navigation parts, and comments. The algorithms will need to be modified upon the changes of page structures of news sources.

## 4 CLUSTERING

Clustering algorithms are similar to those of classification. To find stories that are similar to a particular one, the keyword vector of the story is compared to the keyword vectors of other stories in the database.

## 5 UPDATING

The system's updating schedule is determined by an independent module that can be run on a different server. Upon receiving an updating signal from the scheduler module, the system visits one or several of the 14 different news sources, and then, identifies and retrieves new stories. When and how frequently the system will visit a source are closely related to the source's updating behaviors.

Similar to Web page structures, news-updating behaviors are different for different sources. The updating scheduler is designed as a self-adjusting system that can change its updating schedule when news sources' updating behaviors change. Currently, certain news sources update their news presentations constantly on a 24-hour basis, for example, Agence France Presse, Associated Press, Xinhua News Agency, and especially, Reuters. Other sources' updating is relatively sporadic, for example, The Los Angeles Times's updating occurs mainly at 4:00-7:00AM Eastern time, while USA Today updates its website mainly at 5:00-9:00AM Eastern Time. As a way to reduce network traffic and also to optimize the use of own system resources, those sources that are updated less frequently will also be visited by the system with less frequency.

## 6 CONCLUSION

This system provides users with efficient and reliable access to classified news from different sources. It achieves a high accuracy of classifications with the possibility that one story be classified into more than one category. It also has a high recall of international news from the different news sources.

## REFERENCES


[1] Wen-Lin Hsu and Sheau-Dong Lang. Classification Algorithms for NETNEWS Articles. ACM International Conference on Information and Knowledge Management (CIKM). November 2-6, 1999. Kansas City, Missouri.

[2] Brij Masand, Gordon Linoff and David Waltz. Classifying News Stories using Memory Based Reasoning. 15th Annual International SIGIR, June 1992. Denmark.

[3] Nuno Maria and Mário J. Silva. Theme-based Retrieval of Web News. SIGIR, July 2000. Athens, Greece.

[4] International News Connection, http://www.seventones.com/news/