News Chain Discovery in News Articles*
(Haber Yazılarında Haber Zincirlerinin Keşfi)
TÜBİTAK Project No. 113E249
Fazlı Can (PI), Seyit Koçberber (I)

October 21, 2013

Abstract

Relationships among news articles in a time window form coherent news chains. Most of them are easily recognizable. However, some news chains are not easily discernible due to vague relationships. In topic detection and tracking (TDT) news articles form an easily recognizable news chain. In this study, we aim to discover new chains which are not openly recognizable employing facts not directly seen articles.

When a user sees a news article in a news chain he will be able to see its relationships with the other chain members. If news with different topics are also in the chain the user will be able to realize the hidden relationships among the chained news articles. In this context news chains provide the user seeing the news article in a wide spectrum. In news portals users are overwhelmed by news flood and variety and have difficulties in following their relationships and seeing them in a wide perspective. To overcome this problem we need algorithms that would discover news chains containing the news articles of interest. Google News and similar news portals are typical domains in which the same problem is observed. There is no commercial news portal examples that address this problem. To the best of our knowledge there is no experimental research-oriented example on this subject.

In this study we propose three approaches to discover news chains that include a certain given news article. The first algorithm emphasizes news properties and statistical relationships among news articles. The statistical relationships and similarities among news articles will be respectively computed by employing the language model and vector space model. In the second approach we use social networks that would be generated from news articles in a time window by using a novel algorithm developed within the framework of this project. In the last one these two methods will be used together. In these algorithms we will analyze the conditions related to Turkish such as the effects of stemming and word stopping. The developed methods will be applied to the large scale Bilkent News Portal.

In the course of this study we aim to measure and assess the effectiveness of the developed methods in a reliable way based on scientific user experiments. The observations from the practical portal environment will be reflected to the research process as a feedback, and by that way a synergy will be obtained between research and development.

This project will be conducted by experienced researchers following the principles of the scientific method, which involves measurement, evaluation, and repeatability. The large-scale Turkish news portal that will be enhanced by news chain discovery service will provide effective solutions to real life problems. These features would make it a pioneer in similar applications for Turkish. The nature of the methods that will improve the effective use of news portals and the fact that the characteristics of Turkish will be analyzed in the problem domain add more significance to the project.

The methods to be developed are adaptable to other research and application areas. The news chain discovery can be applied to intelligence applications. For a given intelligence report, intelligence agents can use the discovered chains to determine hidden relationships with other reports. It would make intelligence analysis more efficient and effective. The proposed method can also be used to discover hidden relationships among the documents in huge archives such as blogs and email logs in a similar manner.

* Duration: October 1, 2013 – October 1, 2015

RAs supported by the project: Çağrı Toraman (PhD Student), Tolga Çekiç (MS Student)

Budget 149,044 TL ($73,054 as of October 1, 2013: official project beginning date)