From Newspapers to News Search Systems

Güleda Doğan
Hacettepe University, Department of Information Management, Turkey. Email: guledaduzyol@gmail.com

Abstract: *Today, reading news from internet has become one of the most important reasons for using internet. It is possible to see and read most important news and news topics, also search new and old news with a single search box through news search systems. Main news search systems (news search engines and news metasearch engines) are Google News, AllInOneNews and Yahoo News. This study aims to explain how news search systems work. News clustering and news ranking of news search engines, result extraction component, publication time extraction component, and search engine/database selection component of news metasearch engines will be examined in detail.*

Keywords: *News search systems, news search engines, news metasearch engines.*

Introduction

Main reasons for using news search systems are reading news just as they post, reading new and old news freely which is not possible with newspapers (Liu et al., 2007, p.1017). News search systems covers news search engines and news metasearch engines. The main difference of search/metasearch engines from news search/metasearch engines is that news is time-sensitive. On the other side, main difference between news search engines and news metasearch engines is that news search engines send their web crawlers to news websites periodically, parse the news they crawled, categorize and classify these news (AllInOneNews, 2012; Del Corso, Gullí, and Romani, 2005, p.98; Gulli, 2005, p.880; Liu et al., 2007, p.1017-1018); news metasearch engines transfer user queries to the other search engines, collect news from these search engines and rank these news (AllInOneNews, 2012; Meng, Yu and Liu, 2002, p.54; Liu et al., 2007, p.1017-1018).

News Search Engines

News search engines generally works as Figure 1 (Gulli, 2005).

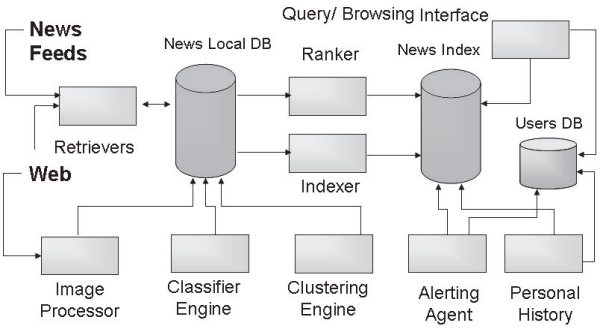

Figure 1. News search engine architecture (Gulli, 2005, p.881)

As Figure 1 indicates, firstly news are retrieved from news sources (via Retrievers - news feeds or web) and stored to a News Local Database, if necessary images are added to the news in News Local Database through Image Processor. News search engines retrieves news from many different news sources Second step is classifying of news in News Local Database using a classifying method/approach. Most of this news comes together with their images, if not (generally, this is the situation for the news retrieved from web) a suitable image for the new is retrieved from the news source and associate to the new in News Local DB with an HTML tag. Each new in News Local DB needs to be classified. Some news have already been classified in the news source, for those that haven't classified different methods are used for classifying and the news associate with a news category such as World, Turkey, Sport, Politics, etc. Next step is measuring similarity of news and clustering of news according to their similarity using different clustering algorithms. Figure 2 illustrates new clustering. According to Figure 2 news from different news sources are clustered according to their similarity using different algorithms. Clustering algorithms used depend on similarity measure used. News clustering can be shown with Gw = S N, where N is news, S is news sources, w is time interval.

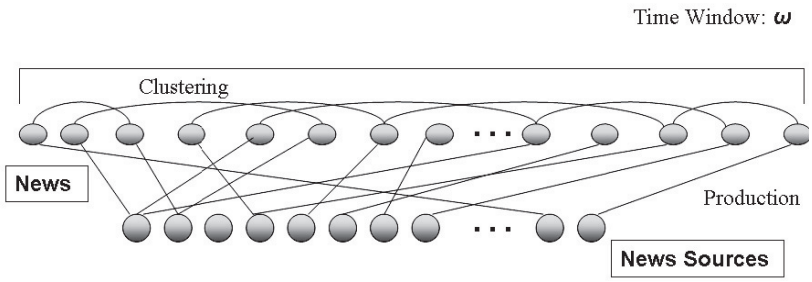

Figure 2. News clustering (Del Corso, Gullí, and Romani, 2005, p. 99)

Ranking of news is very different from ranking of websites because of being time-sensitive. News ranking algorithms ranks news and news sources taking notice of lots of factors such as currency of news, importance of news source etc. After ranking news are indexed. Last step is searching for news through alerting agent, personal history and users databases (AllInOneNews, 2012; Cohen, 2009; Del Corso, Gullí, and Romani, 2005; Gulli, 2005; Newslookup, 2012; Topix, 2012; Wikipedia, 2012).

News Metasearch Engines

General components of news metasearch engines are user interface, search engine connection component, result extraction component, result merging component, search engine/database selection component and publication time extraction component (Liu et al., 2007, p.1017-1018; Meng, Yu and Liu, 2002, p.55). Users type their queries to user interfaces, search engine connection component directs user queries to search engines and bring the result pages through programs, result extraction component extracts search results from result pages, result merging component combines the results (Liu et al., 2007, p.1017-1018; Meng, Yu ve Liu, 2002, p.56).

Search engine/database selection component is necessary for the news metasearch engines contain a large number of search engines (Meng, Yu and Liu, 2002, p.56). A news metaserach engine that contains a large number of news search engines needs an effective "search engine selection algorithm". AllInOneNews uses a revised version of *Optimal Ranking Algorithm.* News search engine selection algorithm determines the best matching news search engines for the user query according to the results brought and uses only these news search engines for the query (Liu et al., 2007, p.1019).

News is one of the most time-sensitive information. Usefullness of current news is the most important reason for this. Current news needs to be ranked upper parts in this context through extracting date and time of news. Publication time extracting component provide extracting the date and time of the news. One of the important problems in terms of publication time extraction is heterogeneity between news sources. For example, meaning of the date 04/03/07 differs by country news come from, April, 3, 2007 or March, 7, 2007. (Liu et al., 2007, p.1019, 1021).

The quality of result merging algorithm directly effects efficiency of news metasearch engine. AllInOneNews uses an improved result merging algorithm. This method determines the rank of each result depend on several factors such as the quality of selected news search engines results come from, common term number between the result title and the query etc. (Liu et al., 2007, p.1021).

Conclusion

Popularity of news search systems are increased with the increasing of reading news from Internet. Almost all newspapers are free from internet today and these online free newspapers take the place of printed newspaper for more people day by day. It is very important to know the algorithm of news search systems for the online newspapers and other news sources to be retrieved and read more.

References

AllInOneNews. (2012). *About us*. http://www.allinonenews.com/aboutUs.html

Cohen, J. (2009). *Same protocol, more options for news publishers*. http://googlenewsblog.blogspot.be/2009/12/same-protocol-more-options-for-news.html

Del Corso, G. M., Gullí, A., & Romani, F. (2005). Ranking a stream of news. In *WWW'05, Special Interest Tracks and Posters of the 14th International Conference on World Wide Web* (p. 880-881), May 10-14, 2005, Chiba, Japan.

Gullí, A. (2005). The anatomy of a news search engine. In *WWW'05, Special interest tracks and posters of the 14th International Conference on World Wide Web* (pp.97-106), May 10-14, 2005, Chiba, Japan.

Bauin, S., & Rothman, H. (1992). Impact of journals as proxies for citation counts. In P. Weingart, R. Sehringer, & M. Winterhager (Eds.), *Representations of science and technology* (pp. 225-239). Leiden: DSWO Press.

Liu, K-L., Meng, W., Qiu, J., Yu, C., Raghavan, V., Wu, Z., Lu, Y., He, H. & Zhao, H. (2007). AllInOneNews: Development and evaluation of a large-scale news metasearch engine. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data* (p.1017-1028), June 11-14, 2007, Beijing, China.

Meng, W., Yu, C., & Liu, K-L. (2002). Building efficient and effective metasearch engines. *ACM Computing Surveys*, *34*(1), 48-84.

Kling, R. & Elliott, M. (1994). *Digital library design for usability*. Retrieved July 19, 2011 from http://www.csdl.tamu.edu/DL94/paper/kling.html

Newslookup. (2012). *About newslookup.com.* http://www.newslookup.com/about.html

Topix. (2012). *About us*. http://www.topix.net/topix/about

Wikipedia. (2012). *Google news*. http://en.wikipedia.org/wiki/Google_News#cite_note-3