# Match the News: a Firefox Extension for Real-Time News Recommendation

Margarita Karkali
Athens University of Economics and Business
karkalimar@aueb.gr

Dimitris Pontikis
Athens University of Economics and Business
pontikhsd@aueb.gr

Michalis Vazirgiannis
Athens University of Economics and Business, LIX - Ecole Polytechnique, Telecom Paris-Tech
mvazirg@aueb.gr

## ABSTRACT

We present *Match the News*, a browser extension for real time news recommendation. Our extension works on the client side to recommend in real time recently published articles that are relevant to the web page the user is currently visiting. *Match the News* is fed from Google News RSS and applies syntactic matching to find the relevant articles. We implement an innovative weighting function to perform the keyword extraction task, $BM25H$. With $BM25H$ we extract keywords not only relevant to currently browsed web page, but also novel with respect to the user's recent browsing history. The *novelty* feature in keyword extraction task results in meaningful news recommendations with regards to the web page the users currently visits. Moreover the extension offers a salient visualization of the terms corresponding to the users recent browsing history making thus the extension a comprehensive tool for real time news recommendation and self assessment.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval** ]: Information filtering

## Keywords

news recommendation, browser extension, novelty detection

## 1. PLUG-IN DESCRIPTION

News recommendation has been a very active area in research and industry in the last several years. This is due to the vast user community and the availability of many news streams. Matching users interests and recommending relevant news is a task that is highly challenging but also quite interesting for the industry as it is many times interleaved with advertising processes. In any case credible news recommendations increase the chances for retaining the user on a specific web page and thus match marketing or other targets. The news recommendation process is based on a user profile that is build based on explicit or implicit data. We simulate users currents interest with the page she is currently visiting and we recommend news relevant to this pages rather than to the generic user profile. This is an innovative news recommendation form to the best of our knowledge. Based on this intuition we designed and built a browser plug-in named *"Match the News"*.


**Figure 1: *Match the News* in action.**

*Match the News* browser extension targets to the average web user and aims at recommending in real time news articles related to the content currently visited. As access to the user's content raises privacy concerns, such a recommendation system should process this information locally or explicitly ask permission from the user to disclose this potentially sensitive information. *Match the News* operates exclusively on the client side and thus respects the user privacy.

The items to recommend are the news articles from the RSS feeds provided by the Google News aggregator. In order to match the web page currently displayed to these articles, keyword extraction on the current web page takes place. The weighting function used for this process is the novel $BM25H$ function introduced in [1]. $BM25H$ takes into account a shifting window of the user's browsing history to construct an dynamic corpus which evolves through time. Using this corpus, we extract keywords that are both relevant to the web page and novel with respect to the previously displayed content. Finally, *Match the News* compares each article to the set of the extracted keywords, and displays the ones having the higher similarity (see figure 1).

*Match the News* has been developed as a Firefox extension and it is available for download from Firefox add-on gallery[1].

[1]https://addons.mozilla.org/addon/match-the-news/


*SIGIR'13,* July 28–August 1, 2013, Dublin, Ireland.
ACM 978-1-4503-2034-4/13/07.

## 2. ARCHITECTURE AND COMPONENTS

**News Articles Collection and Dynamic Corpus.** Implementing a corpus dependent weighting function on a system that works locally on the client's device brings up the problem of corpus storage. Document collections usually used for information retrieval tasks, including the computing of global weights such as $IDF$, present a cardinality ranging in millions, a size prohibitive for a locally operating browser plug-in. *Match the News* develops an evolving corpus consisting of the last N pages the user visits, upon its installation. To effectively maintain such a collection, the plug-in database does not include the actual web documents but only the statistical information for the last N documents, necessary in the $BM25H$ formula, i.e. for each term present in the last $N$ web pages, the $tDF$ value and the time stamp of its last occurrence.

As *Match the News* works on the client side, the items available for recommendation must be locally accessible to select those matching the current user context. This is achieved by periodically collecting the news articles available through Google News RSS feed. Each news unit consists of the article title, a small description (snippet), the URL for the article and the publication date.

**Keyword Extraction.** Real time performance, is a crucial feature for a system that recommends news related to the content *currently* displayed to the user. This requirement, primarily raises from users intolerance to delay, especially for actions not directly related to their current intent. In addition, if the response time of the recommendation system is high, the probability of the user to leave the web page increases, and thus the recommendation process may not conclude. The recommendation process involves matching the web page vector to the respective of the news items. Comparing the full term vector of a web page in order to find the best matching articles can be expensive for large size web pages. To tackle this problem *Match the News* implements a keyword extraction algorithm to select the top-k terms from the web page.

Keyword extraction involves a term weighting function suitable for distinguishing the terms in a web page that better describe the concept presented. *Match the News* implements an innovative weighting function, we introduced in [1], capitalizing on keyword novelty with regards to the keywords distribution within the user's history. This method includes the maintenance of a pseudo-DF, $tDF$, which involves both the document frequency of a term as well as its distribution of occurrences over time and thus declines as the corresponding term stops occurring in the browsing history. Using the $tDF$ to replace the regular document frequency, we introduced $BM25H$, a scoring function that penalizes terms frequently occurring in the browsing history of the user. These terms are often website specific stopwords such as its domain name or other terms globally present in websites, such as menu items (e.g. *contact*, *home* etc). More details and experimental evaluation on the scoring function can be found in [1].

**The News Recommendation Process** When an article reaches the system, it is stored in the database using the bag of words representation. The weighting scheme used for the is the normalized $TF$. Having the term vector for each article and the term vector of the current web page (resulting for the keyword extraction process), *Match the News* uses the cosine similarity measure to compute the similarity between each article having at least one term in common with the web page. From the articles whose similarity to the web page vector passes a certain similarity threshold the top-k are displayed.

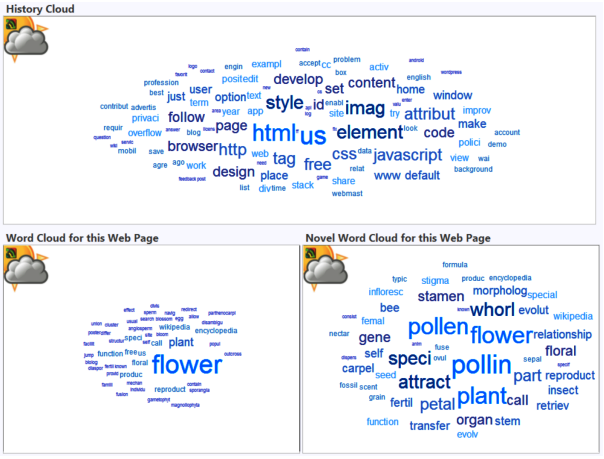


**Figure 2: Word Clouds from *Match the News* for *http://en.wikipedia.org/wiki/Flower*.**

**Recent History Visualization** Alongside the news recommendation service, *Match the News* offers a visualization option, that presents an comprehensive overview of the recent browsing activity and at the same time enables system monitoring by the user herself. Using the popular word-cloud representations, we introduce *History Cloud*, which illustrates the most frequent terms in user's recent browsing history. For the construction of *History Cloud* terms are weighted with $tDF$. Moreover, the current web page is represented by two word clouds. The first is a regular word-cloud, where the terms in the website are weighted using plain $TF$. For the second word-cloud, the terms illustrated are those selected through the keyword extraction process, weighted using $BM25H$. Comparing these clouds we can decide on the representativeness of the corresponding weighting functions. An example of the aforementioned word-clouds can be found in figure 2. There we can easily conclude that the user was recently interested in web development. In addition, comparing the two clouds about the current web page it is obvious that using $BM25H$ instead of plain $TF$ leads to better quality of terms with regards to their discriminative value.

Concluding, we presented a browser extension for real time news recommendation based on the current content the users views. The recommendation process capitalizes on a novel feature extraction measure detecting novel terms in the user browsing history. Moreover it respects user's privacy - as the extension does not keep any centralized user history or other data. The extension offers a salient visualization of the terms corresponding to the users recent browsing history making thus the extension a comprehensive tool for real time news recommendation and self assessment.

## 3. ACKNOWLEDGMENTS

M. Karkali has been co-financed by the EU (ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the NSRF - Heracleitus II.

## 4. REFERENCES

[1] M. Karkali, V. Plachouras, C. Stefanatos, and M. Vazirgiannis. Keeping keywords fresh: a bm25 variation for personalized keyword extraction. TempWeb '12, pages 17–24. ACM, 2012.