Topic Detection and Interest Tracking in a Dynamic Online News Source

Andrew J. Kurtz and Javed Mostafa
Laboratory for Applied Informatics Research
Indiana University, Bloomington
http://lair.indiana.edu/research/newssifter
ajkurtz@indiana.edu, jm@indiana.edu

Abstract

Digital libraries in the news domain may contain frequently updated data. Providing personalized access to such dynamic resources is an important goal. In this paper, we investigate the area of filtering online dynamic news sources based on personal profiles. We experimented with an intelligent news–sifting system that tracks topic development in a dynamic online news source. Vocabulary discovery and clustering are used to expose current news topics. User interest profiles, generated from explicit and implicit feedback are used to customize the news retrieval system's interface.

1. Introduction

With the volume of online news available today, it is difficult to manually sort through the hundreds of daily news articles to find articles related to specific topics. It would be valuable to have an automated system, which would sift through the mounds of news and display the articles that match an individual's interests. Such a system would need to identify topics within the articles, group them into clusters, and present the articles sorted based on the user's interest.

Some advances have been made in filtering, but previous research has generally been in areas that deal with relatively static document sets[4]. Little research has been done in the area of filtering online dynamic news sources based on personal profiles. Some previous research includes Watters and Wang[5] who discuss a system that extracts features from news articles such as date, location, and organization and uses those features to calculate the similarity among articles. SCISOR[3] is a system that analyses news articles to extract the concept of the article. Based on the extracted concepts the articles are summarized and grouped together for use in answering user questions presented to the system.

In addition to filtering news articles into categories, we want to present the articles to the users in a way that focuses their attention on the articles that match their interest. Explicit interest indicators may be used which typically require the user to interrupt their activity and select their interest in a particular topic. This mode is not desirable as it disrupts the news reading process. A better option would be to use implicit interest indicators that are used to gather the users' interest on topics without interrupting the task they are completing. Claypool, et. al.[2] shows that using implicit interest indicators are as accurate in tracking a user's interests as using explicit interest indicators. Explicit interest indicators are a good way for the user to provide their initial interests and implicit interest indicators are good for tracking the user's interest over time.

2. System Design and Methodology

We developed an intelligent news-sifting interface to track topics in a dynamic online news source. Vocabulary discovery and clustering is used to expose current news topics that develop over time. User interest profiles, utilizing both explicit and implicit feedback, track the user's interest and are used to customize the news retrieval system's interface. More information on the algorithms used can be found in Mostafa, et. al.[4]. In this paper, we concentrate on experiments to analyze the impact of key components on system performance.

The system uses an existing online news feed service, called ClariNews[1] that is distributed through USNET newsgroups. We gathered news articles from 384 ClariNews newsgroups that are organized into 25 general interest channels such as "business" and "technology". The update frequency of the channels range between 3 and 300 messages a day. For this paper, we concentrated on the "business" channel that receives an average of 34 messages per day.

We selected three periods covering two weeks to track. All of the articles for the business channel were collected during a total of six weeks. Topic detection and topic clustering were performed on the set of articles. The

Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03)
0-7695-1939-3/03 $17.00 © 2003 IEEE

clusters were analyzed based the number of terms articles were classified into the clusters and the classification was analyzed to see if well-separated clusters were generated.

The evaluation of the interest profile associated with the channels was performed by modeling two types of users. One user was focused on reading one particular news channel for all of the sessions and the second user changed reading habits by switching news channels part way through the sessions. The aim was to see how the interest profiles adapt to each type of user.

3. Results and Analysis

Vocabulary discovery and clustering were performed using components of the SIFTER system[4]. A ranking of tokens based on the tf.idf weights of the tokens was created then the terms in the top R ranks, that appear in at least D documents, were selected. The terms were then clustered using a cosine similarity measure with similarity values below the threshold value Theta being clustered together. The settings of R and D control the number of terms selected and the setting of Theta effects the number of clusters produced. Table 1 shows the cluster results for two settings of R and D and two settings of Theta.

Table 1. Cluster results

	R=20, D=4		R=12, D=2	
	Theta=0.75	Theta=0.50	Theta=0.75	Theta=0.50
Number of Terms				
1/3/03	21	21	65	65
1/17/03	16	16	45	45
1/31/03	13	13	48	48
Period Mean	16.7	16.7	52.7	52.7
Number of Clusters				
1/3/03	7	8	16	28
1/17/03	8	9	15	22
1/31/03	6	8	14	22
Period Mean	7.000	8.333	15.000	24.000
Mean Cluster Distance				
1/3/03	0.974	0.952	0.977	0.957
1/17/03	0.948	0.951	0.964	0.938
1/31/03	0.956	0.911	0.989	0.956
Period Mean	0.959	0.938	0.977	0.950
Mean Cluster Homogeneity				
1/3/03	0.322	0.288	0.429	0.304
1/17/03	0.281	0.259	0.398	0.302
1/31/03	0.316	0.269	0.402	0.256
Period Mean	0.306	0.272	0.410	0.287
Mean Documents Per Cluster				
1/3/03	61.1	54.3	28.8	17.5
1/17/03	80.6	72.5	45.3	31.5
1/31/03	61.9	48.1	28.9	18.8
Period Mean	67.9	58.3	34.3	22.6

discovered and the number of clusters that resulted. The

Providing a larger number of clusters is an advantage to the user since the documents will be separated into more specific topics supporting a higher resolution view of the documents. As we altered R and D to increase the terms selected, thus increasing the number of clusters, the performance of the clustering algorithm did not degrade as observed in the distances between the clusters and the homogeneity of clusters. The results show that we can increase resolution while maintaining performance. In addition, we observed, for both settings of R and D, that increasing the number of clusters improves cluster homogeneity.

The documents were well distributed across the clusters and as the number of clusters increased the mean number of documents in each cluster decreased showing that the new clusters were being created in balance with the document topics.

Documents that did not contain any of the vocabulary terms were classified into a null cluster. As the number of clusters increased the number of documents in the null cluster decreased dramatically, from a mean of 100 (R=20, D=4, Theta=.75) to a mean of 27 (R=12, D=2, Theta=.5), demonstrating that the new clusters were finding new topics within the document set.

User interest levels in the topic channels were modeled using one user with constant on one channel and the second user changing channels part way through. Building profiles according to interest in the channels provides a coarse grain interest profile of the user. A screen shot of the client application can be seen in Figure 1.

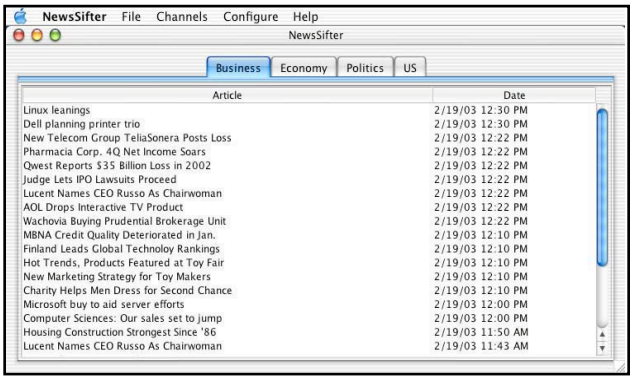

Figure 1. Client interface

Both users initially selected four channels A, B, C, and D and they explicitly set their interest level for each of the channels during the first session at 0.9, 0.8, 0.7, and 0.6 respectively. The level of interest in the four channels was tracked over eight sessions. The interest was tracked using implicit interest indicators.

Figure 2 shows how the interest levels change for a user who monitors and interacts with channel A. The interest level for channel A constantly increases while the interest levels for the other channels constantly decrease.

Figure 3 shows how the interest levels change for a user who focuses on channel A initially then begins to also look at channel B from session five onwards. The change in interest can be seen from the change in the level of channel B as it switches from a downward trend to an upward trend.

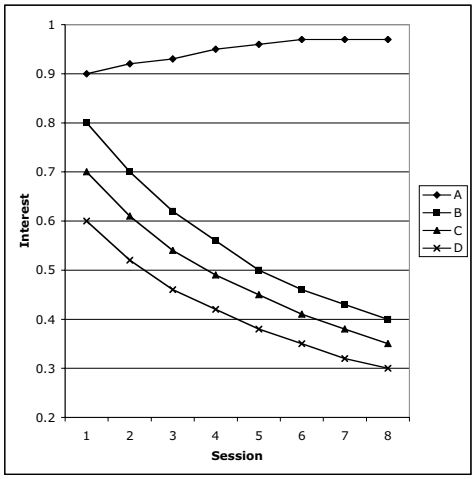

Figure 2. Constant user interest results

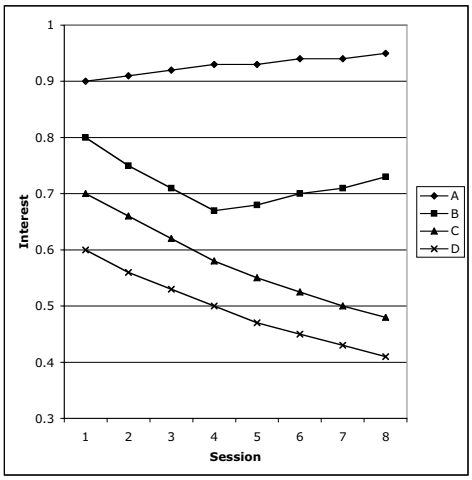

Figure 3. Changing user interest results

As can be seen, the implicit interest indicators used by the system learns a user's interest over time and strengthens the interest level. In addition, the system reacts to a user's change in interest and is able to adapt to the new interest.

4. Conclusion and Future Work

The results described in this paper, show that it is possible to detect topics within a dynamic news source and to track user interest over time. Topic detection and clustering identify the current news topics and a combination of explicit and implicit interest indicators allows the system to track the user's interest with minimal user interruption.

Future work will include adding a second level of interest tracking for the topics within each channel, providing a fine grain interest profile. In addition we will be performing user testing of the system to evaluate performance of the interest tracking based on real users and to evaluate the user interface of the client application.

5. References

[1] ClariNet. http://www.clarinet.com

[2] Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit Interest Indicators. International Conference on Intelligent User Interfaces, 33-40.

[3] Endres-Niggemeyer, B. (1998). SCISOR. In Summarizing Information (pp. 319-327). Berlin: Springer-Verlag.

[4] Mostafa, J., Quiroga, L. M., & Palakal, M. (1998). Filtering Medical Documents Using Automated and Human Classification Methods. JASIS, 49(14), 1304-1318.

[5] Watters, C., & Wang, H. (2000). Rating News Documents for Similarity. JASIS, 51(9), 793-804.