Personalized News Recommendation Based on Click Behavior

在线新闻阅读已经变得非常流行，因为网络提供了从世界各地数以百万计的新闻来源获取新闻文章的机会。新闻网站的一个关键挑战是帮助用户找到那些有趣的文章。本文介绍了在谷歌新闻中开发个性化新闻推荐系统的研究。对于登录并已显式启用web历史的用户，推荐系统根据用户的过去单击行为建立用户新闻兴趣的概要。为了了解用户的新闻兴趣是如何随时间变化的，我们首先对匿名谷歌新闻用户进行了大规模的分析。在日志分析的基础上，我们开发了一个Bayesian框架，用于预测用户当前的新闻兴趣，从特定用户的活动，以及在所有用户活动中显示的新闻趋势。我们将基于内容的推荐机制与现有的协同过滤机制结合起来，生成个性化的新闻推荐。在谷歌新闻中部署了混合推荐系统。对谷歌新闻网站直播的实验表明，这种混合方法提高了新闻推荐的质量，增加了网站流量。

Jiahui Liu, Peter Dolan, Elin Rønby Pedersen
Google Inc.
1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
{jiahui, peterdolan, elinp}@google.com

ABSTRACT
Online news reading has become very popular as the web provides access to news articles from millions of sources around the world. A key challenge of news websites is to help users find the articles that are interesting to read. In this paper, we present our research on developing personalized news recommendation system in Google News. For users who are logged in and have explicitly enabled web history, the recommendation system builds profiles of users' news interests based on their past click behavior. To understand how users' news interests change over time, we first conducted a large-scale analysis of anonymized Google News users click logs. Based on the log analysis, we developed a Bayesian framework for predicting users' current news interests from the activities of that particular user and the news trends demonstrated in the activity of all users. We combine the content-based recommendation mechanism which uses learned user profiles with an existing collaborative filtering mechanism to generate personalized news recommendations. The hybrid recommender system was deployed in Google News. Experiments on the live traffic of Google News website demonstrated that the hybrid method improves the quality of news recommendation and increases traffic to the site.

Author Keywords
Personalization, user modeling, news trend.

ACM Classification Keywords
H.3.3. Information Search and Retrieval: Information filtering.

General Terms
Algorithms, Design, Experimentation

INTRODUCTION
News reading has changed with the advance of the World Wide Web, from the traditional model of news consumption via physical newspaper subscription to access to thousands of sources via the internet. News aggregation websites, like Google News and Yahoo! News, collect news from various sources and provide an aggregate view of news from around the world. A critical problem with news service websites is that the volumes of articles can be overwhelming to the users. The challenge is to help users find news articles that are interesting to read.

Content-based recommendation is a technology in response to this challenge of information overload in general. Based on a profile of user interests and preferences, systems recommend items that may be of interest or value to the user. Content-based methods plays a central role in recommender systems, as it is able to recommend information that has not been rated before and accommodates the individual differences between users [3, 8]. This technique has been applied in various domains, such as email [16], news [4, 5, 20], and web search [15, 18]. In the domain of news, this technology particularly aims at aggregating news articles according to user interests and creating a "personal newspaper" for each user.

基于内容的推荐是应对信息超载的一种技术。根据用户兴趣和偏好的概要，系统推荐可能对用户感兴趣或有价值的项目。基于内容的方法在推荐系统中扮演着核心角色，因为它能够推荐之前没有被评级的信息，并且能够满足用户之间的个体差异[3,8]。该技术已应用于各种领域，如电子邮件[16]、新闻[4,5、20]和网络搜索[15,18]。在新闻领域，这项技术特别针对用户利益聚合新闻文章，为每个用户创建“个人报纸”

An accurate profile of users' current interests is critical for the success of content-based recommendation systems. Some systems [1, 19] require users to manually create and update profiles. This approach places an extra burden on users, something very few are willing to take on. Instead, systems can construct profiles automatically from users' interaction with the system.

In this paper, we describe our research on developing a personalized news recommendation system based on profiles learned from user activity in Google News. The Google News website, available at http://news.google.com, is one of the most popular news websites in the world, receiving millions of page views and clicks from users around the world

The nature of news reading makes news recommendation distinctive from content-based recommendation in other domains. When visiting a news website, the user is looking for *new information*, information that was not known before, or even surprising. Since user profiles are inferred from past user activity, it is important to know how users' news interests change over time and how effective it would be to use the past user activities to predict their future behavior.

To understand this issue, we conducted a large scale log analysis of Google News users to measure the stability of

IUI'10, February 7–10, 2010, Hong Kong, China.

users' news interests. We found that their interests do vary over time but follow the aggregate trend of news events. Based on these findings, we develop a Bayesian model to predict the news interests of an individual user from the activities of that particular user and the news trend demonstrated in activities of a group of users. To recommend news stories to users, the system takes into account of the genuine interests of individual users and the current news trend. Therefore, the user will receive news tailored to her interests without missing the important news events, even when those events do not strictly match the user's particular interests.

We combined the content-based method with the collaborative filtering method previously developed for Google News [7] to generate personalized recommendations for news access. The hybrid method was evaluated in a live experiment: a subset of the live traffic at Google News used the hybrid method; the result showed significant improvement over the existing collaborative filtering method. The experiment on live traffic also revealed a number of interesting issues related to recommendations, serendipitous exploration and user satisfaction. We will discuss these issues later in this paper.

The contribution of this paper is three-fold. First, we report a large-scale log analysis of the consistency of users' news interest. Second, we propose a novel method for predicting a user's news interests based on click behavior which combines the genuine interests of the user and the current news trend. Third, we presented a hybrid system that combines content-based and collaborative filtering method for personalized news recommendation and ran an experiment on the live traffic, showing improved results.

PERSONALIZATION IN GOOGLE NEWS

Google News is a computer-generated news website that aggregates headlines from news sources worldwide. It classifies news articles into different topic categories (e.g. "world", "sport", "entertainment", etc.) and displays them in corresponding sections, as do standard news websites, but with fully automated text-based classification. Google News serves millions of users around the world, and provides numerous editions for different countries and languages.

Users usually visit Google News starting from the homepage. The homepage of the standard edition has the *Top Stories* section on the top of the homepage, followed by topic based sections of news articles, like "world" and "sport".

If a user signs in to her Google Account and explicitly enables Web History, the system will record her click history and generate a personalized section for her, named "Recommended for [*account*]", containing stories recommended based on her click history in Google News. The recorded click histories were fully anonymized and kept secure according to the Google Privacy Policy.

A previous Google News recommendation system was developed using a collaborative filtering method [7]. It recommends news stories that were read by users with similar click history. This method has two major drawbacks in recommending news stories. First, the system cannot recommend stories that have not yet been read by other users, a problem that is often referred to as the *first-rater problem* [7, 8]. For news recommendations, this is a serious problem, as news service websites strive to present the most updated information to users in a timely manner. News articles presented in Google News are usually published within one hour. However, the collaborative filtering method has to wait several hours to collect enough clicks to recommend the news story to users, resulting in undesirable time lags between break-out news and recommendations. Second, not all users are equal to each other, and the collaborative filtering method may not account for the individual variability between users [3]. For example, we observed that entertainment news stories are constantly recommended to most of the users, even for those users who never clicked on entertainment stories. The reason is that entertainment news stories are generally very popular, thus there are always enough clicks on entertainment stories from a user's "neighbors" to make the recommendation.

A solution to these two problems would be to build profiles of user's genuine interests and use them to make news recommendations. The profiles would help the system filter out the stories that are not of interest to the user, such as the entertainment news mentioned above. A news story may also be recommended to the user if it matches her interest, even if the story has not been clicked on by other users.

In this paper, we describe an content-based method to recommend news articles according to their topic categories, which is assigned by text classifiers. Based on a user's news reading history the recommender predicts the topic categories of interest to her each time she visit Google News. News articles in those categories are ranked higher in the candidate list and will be recommended to the user. We chose to recommend news stories at the general level of topic categories instead of fine grained topics because of the nature of news reading: most users visit news websites with the attitude of "show me something interesting," rather than having any specific information goals [7]. Over-specializing the user profile may limit the recommendations to news that the user already knew, which is obviously undesirable for news reading.

The user activities that Google News records are the user's clicks on the Google News website. The system records the event and the time when a user clicks on the page. Each click on a news article is treated as a positive vote for the topic category of that article.

There are two practical constraints on our content-based news recommendation algorithm. First, a user's news interests may change over time. The system should be able to incrementally update the user's profile to reflect change

in interest. Second, there is a large variance in the click history size of the users. A successful algorithm needs to degrade gracefully, i.e. be able to provide reasonable recommendations even when there is little information about the user.

RELATED WORK

Two different technologies are commonly used in recommender systems: content-based recommendation and collaborative filtering. The content-based approach recommends information based on profiles; these profiles are built by analyzing the content of items that the user accessed and favored in the past. In contrast, the collaborative filtering approach does not consider the content of items, but uses the opinions of peer users to generate recommendations. In this paper, we focus on developing effective content-based mechanism for news recommendation in a large-scale website.

The content-based approach has been applied to provide personalized selection of news information in various forms such as personal news agents [1], news readers for wireless devices [2, 3] and web-based news aggregators [19]. These systems build user profiles from information explicitly provided by the user or implicitly observed in user activities. The profiles are then compared with the content of news articles to generate personalized recommendations.

Tan and Teo [19] presented a personalized news system, named PIN. PIN retrieves and ranks news articles according to the user's profile, which is initially defined by the user as a list of keywords and then learned from user feedback using neural network technology. When interacting with PIN, users provide explicit feedback by rating the articles. A similar system, News Dude [1], reads news to users, supporting a series of feedback options such as "interesting", "not interesting", "I already know this", etc. A special purpose news browser for PDAs, named WebClipping2, is implemented by Carreira et al. [3]. WebClipping2 uses a Bayesian Classifier in order to calculate the probability that a specific article would be interesting to the user. Rather than requiring users to provide explicit feedbacks, WebClipping2 observes the total reading time, number of lines read and some other characteristics of user behavior to infer the user's interests. Another personal news agent, PVA [8], uses a proxy to collect user's page clicks and the browsing time, in order to construct a "personal view" that reflects user interests. PVA is applied and evaluated to provide personalized news access.

Unlike these news personalization systems, our news recommendation system infers user interest based on their click behavior on the news website. There are no ratings or negative votes to gauge what the user dislikes. For privacy protection reasons, Google News does not record detailed information about the clicks, such as the amount of time spent on the page. Thus, the system needs to make reasonable prediction with the limited and noisy information of user activity on the website.

Recently, there has been some research on user modeling based on click histories, mostly with the aim of enhancing personalized web search. For instance, Qiu and Cho [13] presented a formal framework and a method to automatically learn user interest based on past click history. The learned user interest is integrated in Topic-Sensitive PageRank to generate personalized ranking. Speretta and Gauch [17] classified queries and snippets of clicked search results to create user profiles, which were then used to re-rank search results. Kim and Chan [9] proposed to model user interest in a hierarchy of concepts, going from general to specific. The hierarchy is learned from the web pages bookmarked by the user using clustering methods.

An important issue in user modeling, particularly for news access, is the changes in user interest over time. Billsus and Pazzani [1] found that there are two types of user interest in news reading: short-term and long-term. The short-term interest usually is related to hot news events and changes quickly. In contrast, long term interest often reflects actual user interest. Accordingly, News Dude [1] uses a multi-strategy machine learning approach to create separate models of short-term and long-term interest. Chen et al. [8] analyzed the change of user interest in news over time and used special mechanisms to update user profiles to reflect user's current interests. Liang and Lai [14] proposed a time-based approach to build user profiles from browsing behavior, which took into account of the time spent by the user on reading the articles and the recency of user activity.

Compared to the above methods, our method is unique in that it captures the dynamic changes of user interest in the context of news trend. The system discovers the genuine interest of users and combines the genuine interest with the current news trend to predict the user's current news interest.

The second technology for recommender systems is collaborative filtering. Collaborative filtering has been applied to personalized news reading applications, such as GroupLens [12] and the first version of Google News recommender.

The content-based and collaborative filtering each has their advantages and limitations [3]. Some research tried to combine both methods and achieved encouraging results [3, 6]. The hybrid method benefits from both methods, providing early predictions that cover all items and users, and improving the recommendations as the number of users and ratings increases. In Google News, we combined our new content-based method with the collaborative filtering method previously developed for Google News [7] to generate personalized recommendations for news access. The live traffic experiment showed that the hybrid method improved the quality of news recommendation.

LOG ANALYSIS OF USER INTEREST 用户兴趣日志分析

The basic assumption of personalization is that users have reasonably consistent interests. A user's history will only be useful if the history help us predict her future actions.

Wedig and Madani [21] performed a large-scale analysis of Yahoo! search engine query logs to determine the consistency of user interests and answer other questions related to personalization. They found that the distribution of users' interest over 22 general topic categories (e.g. "travel", "computing", etc.) converge to a stationary distribution after hundreds of queries are observed. However, news reading is very different from web search. Users usually have specific information goals when issuing queries to search engines, but users visit news website with the attitude of "show me something interesting" [7]. News readers' interest is influenced by the big news events [1]. Many news personalization systems assume that users' news interest change over time [1, 8, 14]. However, to our knowledge, there are no formal studies about how the interest changes. To gain a deep understanding of this issue, we conducted a large-scale log analysis of click behavior on Google News.

Data

We examine the anonymized click logs of those Google News users who were signed into their Google account and explicitly enabled history tracking over 12-month period, from 2007/7/1 to 2008/6/30. From users who made at least 10 clicks per month in that period, we randomly sampled 16,848 users. These users are from more than 10 different countries and regions.

Click Distribution 点击数分布状况

As we described in the previous section, Google News classifies news articles into a predefined set of topic categories, $C = \{c_1, c_2, ..., c_n\}$, including "world", "sports", and "entertainment". In our log analysis, we computed the click distribution over the set of topic categories for individual users as well as the group of users in a country. We divided the time period into 12 months. Then, for each user u, we computed the distribution of her clicks in every month t, $D(u,t)$, represented as a vector over the set of topic categories:

$$D(u,t) = \left(\frac{N_1}{N_{total}}, \frac{N_2}{N_{total}}, ..., \frac{N_n}{N_{total}}\right), \quad \text{where } N_{total} = \sum_i N_i \quad (1)$$

N_i is the number of clicks on articles classified into category c_i made by user u in month t. N_{total} is the total number of clicks made by the user in the time period. Thus, $D(u,t)$ represents the proportion of time the user spent reading about each topic category and reflects the interest distribution of the user in that month.

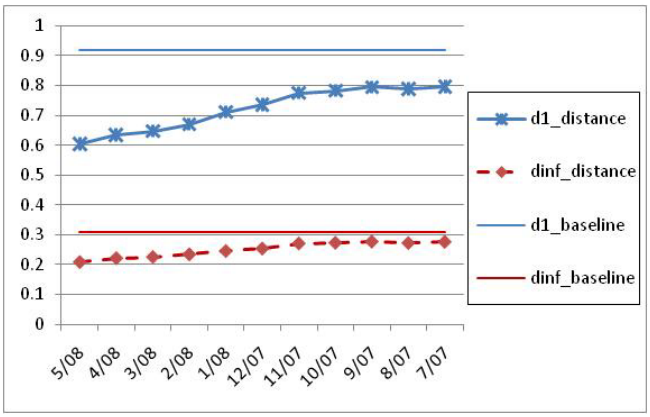

Figure 1. Comparison between the click distribution of the month to be predicted and those of previous months

Change of User's News Interests over Time 随着时间的推移改变用户的新闻兴趣

If a user's news interests are stable, her click distributions in each month should be consistent over time. Particularly, we are interested in using her history to predict her future behavior. Thus, for each user, we compared her click distribution of the most recent month (2008/6/1-2008/6/30) to her click distributions of all the previous months. The comparison demonstrates how well the historical click distribution predicts the future click distribution. Similar to the search log analysis by Wedig and Madani [21], we used d_1 and d_∞, based on l_1 and l_∞, to measure the distance between the click distributions. The d_1 and d_∞ distance of two vectors X and Y is defined as follows:

$$d_1(X,Y) = \sum_i |x_i - y_i| \text{ and } d_\infty(X,Y) = \max_i(|x_i - y_i|) \quad (2)$$

Figure 1 shows the average d_1 and d_∞ distance between the click distribution of the most recent month and those of previous months. Larger value of d_1 and d_∞ distance in a month implies bigger differences between the click distribution of that month and the most recent month. As baselines, we use the d_1 and d_∞ distance of an individual user and the general public in the same location, computed in the next section. As is evident in figure 1, the difference between the click distribution of the most recent month and a past month increased as we go back into the history. Compared to the month of 5/08, the d_1 distance in 7/07 increased 31.9%, and the d_∞ distance increased 32.1%. The figure shows that users' news interests do change over time and their clicks in older history become much less useful in predicting their future interests.

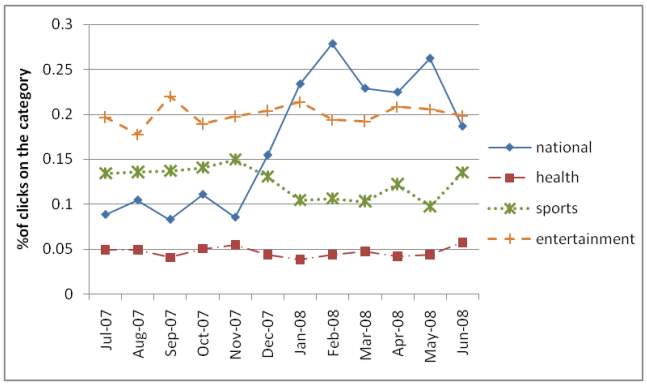

Figure 2. Interest distribution of US users over time

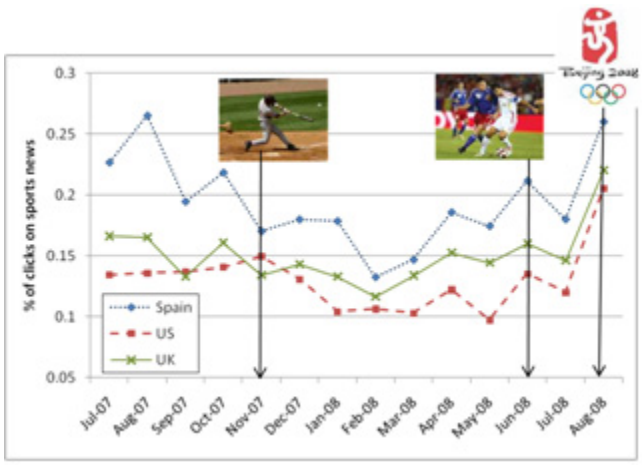

Figure 3. Change of interests in sports news over time

News Trend

In addition to the click distribution of individual users, we calculated the click distributions of the general public in various countries where Google News is served. For each country, the general interests can be represented by the distribution of all the clicks made by the users from that country in a past time period t, represented as $D(t)$.

Figure 2 plots the click distribution for the United States population over time. For the clarity of the figure, only 4 most representative categories are shown in the figure. Figure 2 shows many fluctuations in the news interests of the general public in the US, which was also observed in plots of other countries (not shown in this paper). Furthermore, some topic categories (e.g. "national") showed greater variation than others (e.g. "health"). This phenomenon may be explained by the fact that there are more and bigger break-out news in national politics than health.

We hypothesize that the interest change of a country's general public corresponds to the big news events in that country. The log analysis provided empirical evidence for this hypothesis. For example, the US election campaign starting in late 2007 attracted a large amount of attention to national political news. Figure 2 shows that the percentage of national news clicks doubled during the election campaign compared to before the campaign. Those users who usually paid little attention to national politics probably read more national news about the election campaign because of the importance of the event. Similarly, the 2008 Olympic Games in August 2008 produced a spike in the general interest in sports news in several different countries, as shown in Figure 3.

Moreover, the log analysis shows that there are regional differences in the news trend represented in the click distributions of general public. Figure 3 shows the change of interests in sports news in three different countries: United Kingdom, Spain and United States. Overall, Spanish users read more sports news than British and American users. Figure 3 shows spikes in June 2008 and August 2008, which correspond to the Euro Cup in June and Olympic Games in August respectively. But the American users showed much lower interests in the Euro Cup than the two European countries. On the other hand, the American users' interests in Sports news dropped dramatically in November 2007, when the baseball season ended. However, there were no such trend in Spain and UK.

Influence of the General News Trends on Individual Interest Change 一般新闻趋势对个人利益变化的影响

The previous subsections analyze the interest change of individual users and the general public. A natural question that follows is whether the general news trends influence the interest change of individual users. To understand this question, we compare the click distribution of individual users with the click distribution of the general public in the same time period. We also computed the d_1 and d_∞ distance of an individual user and the general public in a randomly picked different location. If the user's interest is influenced by the local news trend, her click distribution should be more similar to general click distribution of the location that she belongs to than to those of other locations. The average d_1 and d_∞ distance is presented in table 1.

Table 1. Comparison in click distributions between individual users and the general public

	d_1 distance	d_∞ distance
Same location	0.92	0.31
Different location	1.13	0.39

As shown in the table, an individual user's click distribution is more similar to the click distribution of the general public in the same location than to a randomly selected location. Using t-test, both the d_1 and d_∞ distance in the same location are significantly lower than those in the different location, at the confidence level of 99%.

We can draw the following conclusions from this log analysis:

- The news interests of individual users do change over time. 个人用户的新闻兴趣随时间而改变
- The click distributions of the general public reflect the news trend, which correspond to the big news events. 普通大众的点击分布反映了新闻趋势，与大新闻事件相对应
- There exist different news trends in different locations. 不同地区有不同的新闻趋势
- To a certain extent, the individual user's news interests correspond with the news trend in the location that the user belongs to. 在一定程度上，个人用户的新闻利益与用户所属的新闻趋势相符

日志分析显示，个别用户的点击分布受到本地新闻趋势的影响。例如，在欧洲杯期间，西班牙用户阅读更多体育新闻。同样的现象也被报告在一个用户对新闻的生命周期的研究[8]。基于这些发现，我们将用户的新闻兴趣分为两部分:用户的真实兴趣和本地新闻趋势的影响。用户真正的兴趣来自于用户的个人特征，如性别、年龄、职业等，因此随着时间的推移会相对稳定。另一方面，用户在决定阅读内容时，也会受到所在位置的新闻趋势的影响。这种影响会产生短期影响，并随时间而改变。真正的利益和新闻趋势的影响与[1]所讨论的长期和短期利益相对应。然而，我们使用不同的方法来预测用户的新闻兴趣。更重要的是，我们利用公众的点击模式，从新闻趋势的角度对短期利益进行建模，而不是只使用用户自己的反馈。

BAYESIAN FRAMEWORK FOR USER INTEREST PREDICTION 用户兴趣预测的贝叶斯框架

The log analysis reveals that the click distributions of individual users are influenced by the local news trend. For example, Spanish users read more sports news during Euro Cup. Similar phenomena were also reported in a user study of the lifecycle of news interests [8]. Based on these findings, we decompose user's news interests into two parts: users' genuine interests and the influence of local news trend. The user's genuine interests originate from the personal characteristics of the user, such as gender, age, profession, etc. and are thus relatively stable over time. On the other hand, when deciding what to read, users are also influenced by the news trend in the location that they belong to. This kind of influence produces short-term effects and changes over time. The genuine interests and news trend influence correspond to the "long-term" and "short-term" interests discussed in [1]. However, we used distinct methods to predict user's news interests. More importantly, we model the "short-term" interests from the perspective of news trend using the click patterns of the general public, instead of only using the user's own feedbacks.

We developed an approach using Bayesian frameworks [10] to predict users' current news interest based on the click patterns of the individual users and the group of users in the country. The predicted interests are used in news recommendation. The approach works as follows: first, the system predicts user's genuine news interests regardless of the news trend, using the user's clicks in each past time period; second, the predictions made with data in a series of past time periods are combined to gain an accurate prediction of the user's genuine news interests; finally, the system predicts the user's current interests by combining her genuine news interests and the current news trend in her location.

Predicting User's Genuine News Interest 预测用户真正的新闻兴趣

For a specific time period t in the past, we observed the click distribution of individual users, $D(u,t)$, and the click distribution of all the users in a country, $D(t)$, which represents the news trend in that country in the time period. We would like to learn the user's genuine interests revealed in $D(u,t)$ regardless of the influence of $D(t)$. The genuine interest of a user in topic category c_i is modeled as $p^t(click \mid category = c_i)$, the probability of the user clicking on an article about c_i. Using a Bayesian rule, $p\ (click \mid category = c_i)$ is computed as follows:

$$interest^t(category = c_i) = p^t(click \mid category = c_i) = \frac{p^t(category = c_i \mid click)\, p^t(click)}{p^t(category = c_i)} \quad (3)$$

$p^t(category = c_i \mid click)$ is the probability that the user's clicks being in category c_i. It can be estimated by the click distribution $D(u,t)$ observed in time period t, as computed in Equation 1.

$p^t(category = c_i)$ is the prior probability of an article being about category c_i. This is the proportion of news articles published about that category in the time period, which correlates with the news trend in the location. As more news events happen in a given topic category, more news articles will be written in that category. Thus, we can approximate this probability with the click distribution of the general public $D(t)$.

$p^t(click)$ is the prior probability of the user clicking on any news article, regardless of the article category.

According to Equation 3, $p(click \mid category = c_i)$ represents the extent to which the user's interest in the topic category differs from the general public of the same location. If the user reads a lot of sports news while a lot of users are reading it, the user may not be particularly interested in sports but read the sports news because of some hot sports event. In contrast, an extraordinary large proportion of clicks on sports news is a strong signal for the user's genuine interests in sports.

Combining Predictions of Past Time Periods 结合过去时期的预测

Equation 3 computes the user's genuine news interest based on the click distributions in a particular time period. To accurately gauge the user's genuine interests, we combine the predictions made over multiple time periods as follows:

$$interest(category = c_i) = \frac{\sum_t \left(N^t \times interest^t(category = c_i)\right)}{\sum_t N^t} = \frac{\sum_t \left(N^t \times \frac{p^t(category = c_i \mid click)\, p^t(click)}{p^t(category = c_i)}\right)}{\sum_t N^t} \quad (4)$$

The more clicks we have recorded about the user, the better the prediction is going to be. Therefore, we normalize the

predictions made in time periods t by N^t, the total number of clicks by the user in time period t.

We can assume that the prior probability of a user clicking on any article is constant over time. Thus, Equation 4 becomes Equation 5:

$$interest(category = c_i) = \frac{p(click) \times \sum_t \left(N^t \times \frac{p^t(category = c_i \mid click)}{p^t(category = c_i)} \right)}{\sum_t N^t} \quad (5)$$

Predicting User's Current News Interest 预测用户当前的新闻兴趣

正如我们之前讨论的，用户的新闻兴趣被分解为两个部分：真正的新闻兴趣和新闻趋势的影响。上一节基于她的过去点击行为计算用户真实的新闻兴趣。为了判断当前的新闻趋势，我们在短时间内使用普通大众的点击分布

As we discussed before, the user's news interest is decomposed into two parts: the genuine news interest and the influence of news trends. The previous section calculated the user's genuine news interests based on her past click behaviors. To gauge the current news trend, we use the click distribution of the general public in a short current time period (e.g. in the past hour), represented as $p^0(category = c_i)$. Because of the large number of users, there are enough clicks in the short current time period to accurately estimate the popular topic categories in the location.

The ultimate goal is to predict the click distribution of the user for the near future. Again, we use the Bayesian law:

$$p^0(category = c_i \mid click) = \frac{p^0(click \mid category = c_i) p^0(category = c_i)}{p^0(click)} \quad (6)$$

We estimate $p^0(click \mid category = c_i)$ with the genuine news interests, $interest(category = c_i)$, computed in Equation 5, and assume the probability a user clicking on any news article is constant, thus,

$$p^0(category = c_i \mid click) \propto \frac{interest(category = c_i) p^0(category = c_i)}{p(click)} \propto \frac{p^0(category = c_i) \times \sum_t \left(N^t \times \frac{p^t(category = c_i \mid click)}{p^t(category = c_i)} \right)}{\sum_t N^t} \quad (7)$$

In addition to the user's past clicks, we add a set of virtual clicks, with the same click distribution as that of current news trend, i.e. $p^0(category = c_i)$. Thus, the final estimation of the user's news interests in the near future is

$$p^0(category = c_i \mid click) \propto \frac{p^0(category = c_i) \times \left(\sum_t \left(N^t \times \frac{p^t(category = c_i \mid click)}{p^t(category = c_i)} \right) + G \right)}{\sum_t N^t + G} \quad (8)$$

G is the number of virtual clicks (set to be 10 in the system), which can be regarded as a smoothing factor. When the system observes very few (even zero) clicks from the user, the system will predict the user's interest mostly based on the current news trend, which is still a reasonable estimation. On the other hand, if $\sum_t N^t$ is the much larger than G, the estimation is mainly based on the user's own click distribution in the past.

Another advantage of the proposed approach is that the user's interests can be updated incrementally. The system can save the values of N^t and $\frac{p^t(category = c_i \mid click)}{p^t(category = c_i)}$ for each past time period. When updating the user's profile, the system only needs to compute the value for the most recent time period and recompute the weighted sum with the saved values.

NEWS RECOMMENDATION

In order to rank the list of candidate articles to be recommended, the system generates an content-based recommendation score, $CR(article)$, and a collaborative filtering score, $CF(article)$, for each article. $CR(article)$ is based on the topic category of that article and the predicted user's interest using Equation 8. The collaborative method implemented in [7] computes $CF(article)$. The two scores are combined in ranking the candidates for news recommendation:

$$Rec(article) = CR(article) \times CF(article) \quad (9)$$

Combining the content-based method and the collaborative method offers the advantages of both methods and shows improved performance over using the collaborative method alone. In the next section, we describe our evaluation of the hybrid method on the live traffic of Google News.

LIVE TRAFFIC EXPERIMENT

To evaluate the performance of the hybrid methods and understand the user experience with personalized news recommendation, we conducted experiments on a fraction (about 10,000 users) of the live traffic at Google News. The users were randomly assigned to a control group and a test group. The two groups had about the same number of users. When a logged-in Google News user (who also explicitly has enabled web history) visits the website, a section of recommended news is generated particularly for that user. In our experiment, the users in the control group get recommended news from the existing collaborative filtering

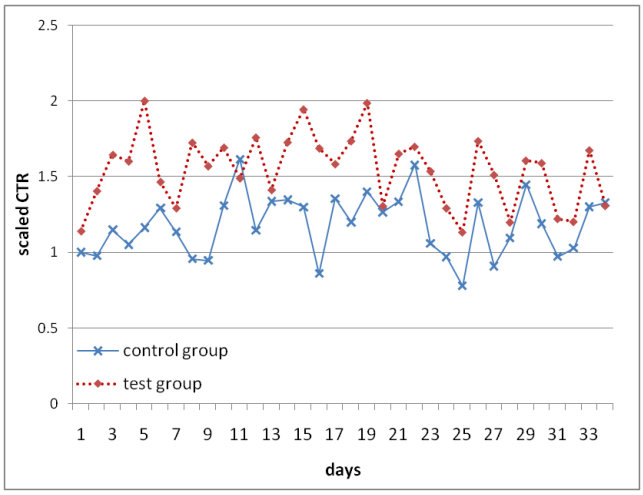

Figure 4. CTR of the recommended news section

method; while the new hybrid method is used for the test group. Aggregate click-through rate analysis was then performed over fully anonymized click logs.

The experiment was run for 34 days, from 1/10/2009 to 2/17/2009. The user's clicks in the past 12 months are used as history to compute the user's interests. To gain greater accuracy in estimating the news trend in the past, we calculated the click distributions of the general public for each week. The current news trend, $p^0(category = c_i)$, is estimated with the click distribution of the general public in the past day.

Three different metrics are used to measure the performance of the recommender and the user's experience: click-through rates (CTR) of the recommended news section, CTR of the Google News homepage, and frequency of visiting Google News website. We calculated the three metrics for each user on daily basis. The performance of the control and test group was derived by averaging the measurements of all the users in the corresponding group. We report the experiment results for the three aspects below.

CTR of the recommended news section is calculated as the number of clicks on the recommended news articles every time the user visits the Google News website. It directly measures the quality of the recommendations as how many of the recommendations are clicked on, thus liked, by the user. Figure 4 shows the CTR of the recommended news section for the control and test group in the 34 days. The values are scaled so that the CTR of the control group in the first day is 1. As shown in the figure, the CTR in the test group is consistently higher than the CTR in the control group, in 33 of 34 days in the experiment. This shows that the proposed news interest prediction method improved the quality of news recommendations. On average, the hybrid method that incorporates the information filtering method improves the CTR upon the existing collaborative method

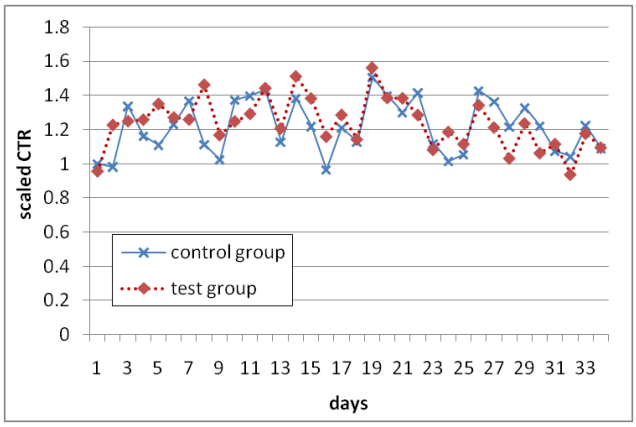

Figure 5. CTR of the Google News homepage

by 30.9%. The improvement is significant at the 99% confidence level according to t-test.

The recommended news section is only one part of the Google News website, which presents to the user many other standard non-personalized news sections along with the recommended news section, such as top stories, world news, business news, etc. We would like to analyze the effect of the improved recommender on the user experience of the whole website. Two metrics are computed to evaluate the news recommender in the larger context of news reading: CTR of the Google News homepage and the frequency of visiting the Google News site.

The CTR of the Google News homepage is calculated as the total number of clicks for each page visit made by the user. Figure 5 plots the measurements for the control and test groups in the experiment. Interestingly, there is not much difference in the CTR of the homepage for the two groups. Although the test group clicked on more news articles in the recommended news section (shown in figure 4), the total number of articles that a user is willing to click on in each website visit seems to be constant. In other words, the improved recommender "stole" clicks from other non-personalized sections, rather than increasing the overall number of clicks. The experiment demonstrated that the

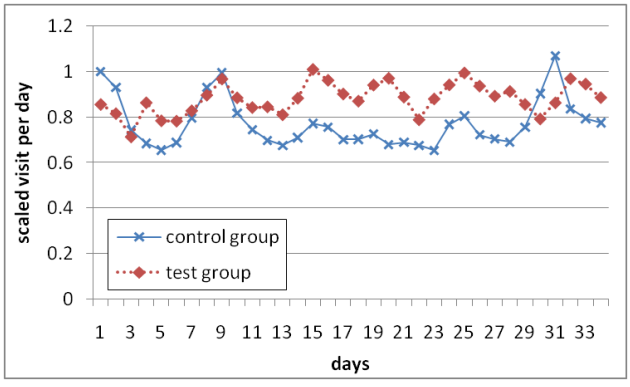

Figure 6. Frequency of website visit per day

improved news recommender created a more focused news reading in the test group. As the recommender was improved to present news articles that better matched the user's interests, the users seemed to pay more attention to the recommended news section and spend less time and effort in finding interesting news articles in the non-personalized sections.

We measure the overall satisfaction of the Google News website with the frequency of website visits, calculated as the number of times the user visits the website in a day. Figure 6 shows the frequency of website visit for the control and test group. It is evident in the figure that the test group visited Google News more often than the control group in most of the days in the experiment period. On average, the frequency of website visits in the test group is 14.1% higher than the control group. The improvement is significant at the 99% confidence level according to t-test.

In summation, the proposed news interest prediction method improved the quality of news recommendations. More recommended news articles were clicked on by the users in the test group using the new hybrid method than the control group using the existing collaborative filtering method. As a result, users seemed to like Google News more and visited the website more often. However, the total amount of attention that users are willing to pay per visit seems to be constant. As users clicked on more recommended news articles, they clicked on fewer articles in the standard non-personalized sections. More research of in-depth user studies would be needed to understand the effects of personalization on information exploration and serendipitous discovery.

CONCLUSION AND FUTURE WORK

In this paper, we present our research on developing an effective information filtering mechanism for news recommendations in a large-scale website such as Google News. We first conducted a log analysis on the change of user's interests in news topics over time. The log analysis demonstrated variations in users' news interests and shows that the news interests of individual users are influenced by the local news trend. Based on these findings, we decompose users' news interests into two parts: the genuine interests and the influence of local news trends. A Bayesian framework is proposed to model a user's genuine interests using her past click history and predict her current interests by combining her genuine interest and the local news trend. The method for predicting user's interests was used in content-based news recommendation, and it was combined with the existing collaborative filtering method to generate personalized news recommendations. We conducted an experiment with the news recommender using the hybrid method on a fraction of live traffic on the Google News website. Compared with the existing collaborative filtering method, the experiment showed that the hybrid method improved the quality of news recommendations and attracted more frequent visits to the Google News website.

The research can be extended in the following directions in the future. Position bias can be investigated and incorporated in modeling users' interests using the click behavior. More advanced methods for combining the information filtering and collaborative filtering mechanisms can also be studied to better leverage the advantages of both mechanism. In addition, our live traffic experiment revealed that the improved recommender increased the CTR of the recommended news sections while reducing the CTR of other standard sections. Further user studies can be conducted to investigate this phenomenon to better understand the effect of personalization on news exploration.

REFERENCE

1. Billsus, D. and Pazzani, M. J. A hybrid user model for news story classification. In Proceedings of the Seventh International Conference on User Modeling. 1999.
2. Billsus, D. and Pazzani, M. J., User Modeling for Adaptive News Access, User Modeling and User-Adapted Interaction, v.10 n.2-3, p.147-180, 2000
3. Carreira, R., Crato, J. M., Gonçalves, D., Jorge, J. A. Evaluating adaptive user profiles for news classification, Proceedings of the 9th international conference on Intelligent user interfaces, 2004.
4. Chen, C. C., Chen, M. C., Sun, Y. PVA: a self-adaptive personal view agent system, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001.
5. Chen, Y-S., Shahabi, C.: Automatically improving the accuracy of user profiles with genetic algorithm. In: Proceedings of IASTED International Conference on Artificial Intelligence and Soft Computing, 2001.
6. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. and Sartin, M. Combining Content-Based and Collaborative Filters in an Online Newspaper. In Proceedings of ACM SIGIR Workshop on Recommender Systems, 1999.
7. Das, A. S., Datar, M., Garg, A., Rajaram, S. Google news personalization: scalable online collaborative filtering, Proceedings of the 16th international conference on World Wide Web, 2007
8. Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J. Combining collaborative filtering with personal agents for better recommendations, Proceedings of the 16th national conference on Artificial intelligence and the 11th Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, 1999.
9. Kim, H. R., Chan, P. K. Learning implicit user interest hierarchy for context in personalization, Proceedings of

the 8th international conference on Intelligent user interfaces, January 12-15, 2003.

10. Jensen, V. Bayesian Networks and Decision Graphs. Springer, 2001

11. Katakis, I., Tsoumakas, G., Banos, E., Bassiliades, N., Vlahavas, I. An adaptive personalized news dissemination system. In Journal of Intelligent Information Systems, Volume 32, Issue 2. 2009.

12. Konstan, J. A., Miller, B.N., Maltz,D., Herlocker, J. L.,Gordon, L. R., and Riedl, J. Group-Lens: Applying collaborative filtering to usenet news. Commun. ACM 40, 77-87. 1997.

13. Lee, U., Liu, Z., Cho, J. Automatic identification of user goals in Web search, Proceedings of the 14th international conference on World Wide Web, 2005

14. Liang, T.-P. and Lai, H.-J. Discovering User Interests from Web Browsing Behavior: An Application to Internet News Services, IEEE Computer Society, Los Alamitos, CA, USA, 2002.

15. Liu, F., Yu, C., Meng, W. Personalized Web Search For Improving Retrieval Effectiveness. In: IEEE Transactions on Knowledge and Data Engineering, 2004.

16. Maes, P. Agents that reduce work and information overload, Communications of the ACM, v.37 n.7, p.30-40, July 1994.

17. Speretta, M., Gauch, S.: Personalized Search based on User Search Histories. In: IEEE/WIC/ACM International Conference on Web Intelligence, 2005.

18. Sugiyama, K., Hatano, K., Yoshikawa, M. Adaptive web search based on user profile constructed without any effort from users. In: Proceedings 13th International Conference on World Wide Web, 2004.

19. Tan, A. and Tee, C. "Learning User Profiles for Personalized Information Dissemination," Proceedings of 1998 IEEE International Joint conference on Neural Networks, pp. 183- 188, May 1998

20. Tan, A., Teo, C.: Learning user profiles for personalized information dissemination. In: Proceedings of 1998 IEEE International Joint Conference on Neural Networks, 1998.

21. Wedig, S., Madani, O. A large-scale analysis of query logs for assessing personalization opportunities, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006.