Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2012 Proceedings

Proceedings

# Analysis of Probabilistic News Recommender Systems

Shankar Prawesh
*Information Systems and Decision Systems, University of South Florida, Tampa, FL, United States.*, shankar1@usf.edu

Balaji Padmanabhan
*Information Systems and Decision Systems, University of South Florida, Tampa, FL, United States.*, bp@usf.edu

## Recommended Citation

Prawesh, Shankar and Padmanabhan, Balaji, "Analysis of Probabilistic News Recommender Systems" (2012). *AMCIS 2012 Proceedings*. 5.
http://aisel.aisnet.org/amcis2012/proceedings/DataInfoQuality/5

# Analysis of Probabilistic News Recommender Systems

**Shankar Prawesh and Balaji Padmanabhan**
Information Systems and Decision Sciences
College of Business, University of South Florida
4202 E. Fowler Avenue, Tampa, FL 33620
{shankar1,bp}@usf.edu

## ABSTRACT

The focus of this research is the *N* "most popular" (Top-N) news recommender systems (NRS), widely used by media sites (e.g. New York Times, BBC, Wall Street Journal all prominently use this). This common recommendation process is known to have major limitations in terms of creating artificial amplification in the counts of recommended articles and that it is easily susceptible to manipulation. To address these issues, probabilistic NRS has been introduced. One drawback of the probabilistic recommendations is that it potentially chooses articles to recommend that might not be in the current "best" list. However, the probabilistic selection of news articles is highly robust towards common manipulation strategies. This paper compares the two variants of NRS (Top-N and $probabilistic$) based on (1) accuracy loss (2) distortion in counts of articles due to NRS and (3) comparison of probabilistic NRS with an adapted $influence\ limiter$ heuristic.

### Keywords

Kullback-Leibler distortion, news recommender systems, influence limiter, accuracy loss.

## INTRODUCTION

In the last ten years the Web has grown to become the primary news source for many users. At the same time there has been bigger penetration of social media (such as tweets, Facebook posts and amateur online videos) (The Economist, 2011). The change in news consumption behavior of readers has made the news industry more participatory where readers often volunteer to submit, share and comment on news articles. As the Economist writes, "*the most popular stories cause a flood of traffic as recommendations ripple across social networks*" (The Economist, 2011). Hence, once an article makes it into a "most popular" list of NRS, there can be a self-reinforcing effect that can further impact its ultimate readership or influence.

The focus of the present research is to discuss the tradeoff between Top-N and probabilistic NRS in terms of count distortion and information quality[1]. Top-N NRS uses a "hard cutoff" and selects the *N* articles with the highest counts to be displayed as recommendation; whereas in probabilistic NRS, recommendations are generated with a probabilistic sampling without replacement of *N* articles. The probability that an article 'a' gets recommended in probabilistic NRS at time *t* depends on the current count of this article and the counts of all other articles. Further, the recommendation probability is proportional to the count of an article at time *t*. The precise mathematical expression for probabilistic NRS is discussed later in the simulation model.

Findings of present research have been presented through a thought experiment implemented as a simulation. In the Top-N NRS the $(N+1)^{th}$ article, which may have "just" missed the cutoff, is often unduly penalized in terms of readership counts in the long run. This weakness of the Top-N recommender can be exploited by manipulators who seek to gain popularity for their articles. We show that the probabilistic mechanism is more robust towards common manipulation strategies. Finally, we present results on manipulation for the probabilistic NRS in comparison with an "adapted" influence limiter heuristic (Resnick and Sami, 2007).

## RELATED WORK

In one of the earliest research in online manipulation, Dellarocas has discussed the theoretical analysis of manipulation strategies and its impact on the firm and consumer in a simple monopoly setting (Dellarocas, 2006). In a related work on manipulation Resnick et al. introduced the influence limiter algorithm for items recommendation, controlling rater's influence on recommender systems through reputation acquired over time (Resnick et al., 2007). In subsequent work (Resnick et al., 2008) established the tradeoff between resistance to manipulation by an attacker and optimal use of genuine ratings in recommender systems. A lower bound on how much information must be discarded is also provided. In a slightly

[1] Counts of articles has been assumed as the surrogate measure of quality

different approach $Shilling\ attacks$ detection on recommender systems has been proposed as a two phase procedure. First, a multidimensional scaling has been used to identify distinct behavior and to narrow down detection space by filtering out noise profiles. In the second phase a clustering based method has been used to discriminate the attackers (Lee and Zhu, 2011). In more recent work the existence of an "*Internet Water Army*" who get paid for posting comments, threads and news articles has been discussed. These groups are known to "flood" the internet with purposeful comments and articles (Chen, Wu, Srinivasan and Zheng, 2011). Techniques to identify such manipulators from behavioral and semantic data and has been also discussed (Chen et al., 2011).

In particular for NRS, a robust voting system for social news websites based on SpotRank has been introduced. Considering voting as a recommendation, Lergillier et al. have presented a set of heuristics that demotes the effects of manipulation. SpotRank is built over $ad-hoc$ statistical filters, a collusion detection mechanism and also the reputation of users and proposed news (Lergillier, Peyronnet and Peyronnet, 2010). Analytical modeling for the news aggregation process by Digg for news recommendation and ratings has been also addressed (Lerman, 2007). Finally, limitations of popularity based ranking are also discussed in context of web ranking (Cho, Roy and Adams, 2005). To alleviate the bias related with popularity based ranking function based on *page quality* has been introduced.

## MODEL

To compare the two variants of the NRS, we set up a simulation model as follows. We maintain a *Comprehensive List* (CL) of articles and their counts. From $CL$, we "select" (described below) $N$ articles into the *Display List* (*DL*) as "recommendations". Initially articles are assigned random counts. At the start of the simulation the $(N+1)^{th}$ article was deliberately assigned a count of exactly one less than the count of $N^{th}$ article to study emergent behavior from just small differences initially.

*DL* is updated periodically and is based on two selection processes - *count-based* (i.e. Top-N) and *probabilistic selection*. As mentioned earlier, count-based selection uses a "hard cutoff" and selects the *N* articles with the highest counts. This is how most online news sites prominently display the most popular or viewed articles. *Probabilistic selection* is a mechanism where every article in *CL* will have some probability, based on its count, to appear in *DL*. The probability that an article will be selected in $DL$ is given by $prob_a(t) = \frac{count_a(t)}{\sum_j count_{j}(t)}$, where $count_a(t)$ represents the count of an article '*a*' at a given time *t* and $\sum_j count_j\ (t)$ represents the total counts of articles those are not yet selected for *DL* at time *t*. This sampling is repeated *N* times without replacement to generate the *N* recommendations in *DL*.

Two different reader models were also implemented. In both a user selects an article from *DL* with probability $p$ or from *RL* (=*CL*-*DL*) with probability $1 - p$. In the first user model a reader selects an article from *DL* randomly. In the second model the top article in the *DL* has the highest probability of being selected and the last has the lowest probability with a linear decrease in the selection probability. Specifically, for the second reader model the probability of a particular article with rank $i, i \in \{1,2,..N\}$ in $DL$ being read (selected) is given by $r_i = \frac{N+1-i}{\sum_{i=1}^{N} i}$. For ease of exposition the present model intentionally leaves out other factors of news arrival and reader behavior.

The pseudo code for the simulation is presented below ("*Select*" can be count-based or probabilistic; while "*Choose*" can be based on either of the two reader models described above).

```
For each reader
    Sort the updated count and select $N$ articles for $DL$
    If selected article is from $DL$(i.e with probability $p_1$)
        Choose an article from $DL$ and Increase its count by 1
    Else
        Randomly choose an article from $RL;(RL = CL - DL)$.and increase its count by 1
End for.
```

In order to compare the effect of manipulation on different selection mechanisms we consider one specific measure here. This measure is based on the counts of $N^{th}$ and $(N+1)^{th}$ articles over the complete simulation. Both $N^{th}$ and $(N+1)^{th}$ articles selected here are based on the initial counts of articles before the simulation starts. Measure *M1* is defined as the logarithmic-ratio of the counts of $N^{th}$ and $(N+1)^{th}$ articles as follows:

$M1(t) = \ln\frac{count_{Nt}}{count_{(N+1)t}}$ measured at time $t$ in the simulation. This measures the relative change in counts of $N^{th}$ and $(N+1)^{th}$ article. At the start of simulation, $count(N) \sim count(N+1)$, hence $M1(0) \sim 0$.

We assume that a manipulator can create artificial clicks to raise the counts of a selected article (such as by creating fake IDs for instance). The particular article selected for manipulation in the present model is the $(N+1)^{th}$, since this is the article that would have just missed a hard "top$-N$" cutoff. The focus of this research is manipulation at an early stage. In "early" manipulation, the fake clicks are assumed to be distributed in some early part of the time period. We also examine the performance of both probabilistic NRS and influence limiter algorithm (for the Top-N NRS) for the extent of manipulation (high and low, based on how many fake counts are generated). The parameters used in the simulation are given in table 1.

| **Parameter** | **Value** |
|---|---|
| Number of readers | 1500 |
| Number of articles in *DL* | 10 |
| Number of articles in *CL* | 200 |
| Initial counts of articles[2] | Random integer between 0 and 1000 |
| Manipulation counts | 10 and 50 |
| Probability of selection of an article from *DL* | 0.9, 0.5, 0.25, 0.1 |

**Table 1: The Model Parameters Used in the Simulation**

To get maximum "benefit" early manipulation is very effective when Top-N NRS is used. Because through injecting artificial clicks on the early part of a news article's lifespan manipulators can leverage the self-reinforcing nature of the count-based Top-N recommender to work in their favor to gain more popularity. These results, demonstrating the advantages of probabilistic NRS, have been shown in prior work (Prawesh and Padmanabhan, 2011). Extending these ideas, this paper examines properties of the probabilistic NRS in greater detail.

## ANALYSIS OF PROBABILISTIC NRS

In this section we analyze probabilistic NRS in two ways. First we present and discuss an accuracy-distortion tradeoff. Then we compare it against a novel adaptation of the influence limiter algorithm.

### Comparison of NRS – The Accuracy/Distortion Tradeoff

*Accuracy (MAE)*

One drawback of the probabilistic recommendations is that it potentially chooses articles to recommend that might not be in the current "best" list. To quantify that loss in the recommendation process, the Top-N and probabilistic NRS are compared based on the "quality" (measured as popularity) of the articles appearing in the recommended list. A widely used measure for this purpose is mean absolute error (MAE). It represents an efficient means to measure the statistical accuracy of predictions of articles appearing in the Top-$N$ recommendation (Ziegler, McNee, Konstan and Lausen, 2005). Let us denote the count of articles appearing in count based NRS as $N_h$ and probabilistic NRS as $N_p$. The metric denoted as $|\bar{E}|$ is given by,

$$|\bar{E}| = \frac{1}{|j|}\sum_{j}\frac{|\sum_i N_h(i) - \sum_i N_p\,(i)|}{|\sum_i N_h\,(i)|}$$

The summation is over all the articles appearing in the recommended list at any given time; $j$ represents the number of time steps in the simulation. MAE presents accuracy in terms of "high" ranked articles assuming that users will have little or no interest in the "low" ranked articles, averaged over the complete simulation.

*Distortion (KL)*

Assuming that the initial share of articles represents the "true" preference of readers, the distortion created by each NRS in comparison with their initial share is given by $Kullback - Leibler$ $(KL)$ distortion measure (Kullback and Leibler, 1951).

[2] Except $N^{th}$and $(N+1)^{th}$ articles. Counts for these articles were assigned such that $count(N+1) = count(N) - 1$. This was done deliberately to test the impact of little early manipulation.

Let us denote the probability distribution for articles in each NRS (probabilistic and Top-N NRS) at the time $t$ as $q_t(x_i)$. Then the $KL$ distortion for the articles $\{x_1, x_2, \ldots \ldots x_n\}$ is given by.

$$D_{KL}(p\|q_t) = \sum_{i=1}^{n} p(x_i) \ln \left(\frac{p(x_i)}{q_t(x_i)}\right)$$

In other words, the above expression represents the inefficiency of the distribution $q$ when the true distribution of articles is $p$ (given initially).

Since, the emergence of counts of the articles in a given NRS is probabilistic process; the data was generated through fifteen replications of the complete simulation for the different values of reading probability for both reader models. The results discussed below are based on the mean value of metric over fifteen replications, plotted against different choice of reading probabilities.

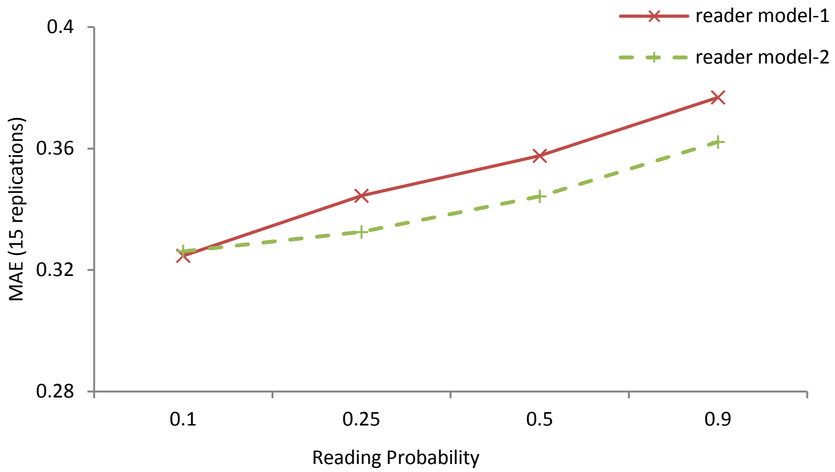

**Figure 1: Mean Absolute Error vs. Reading Probability**

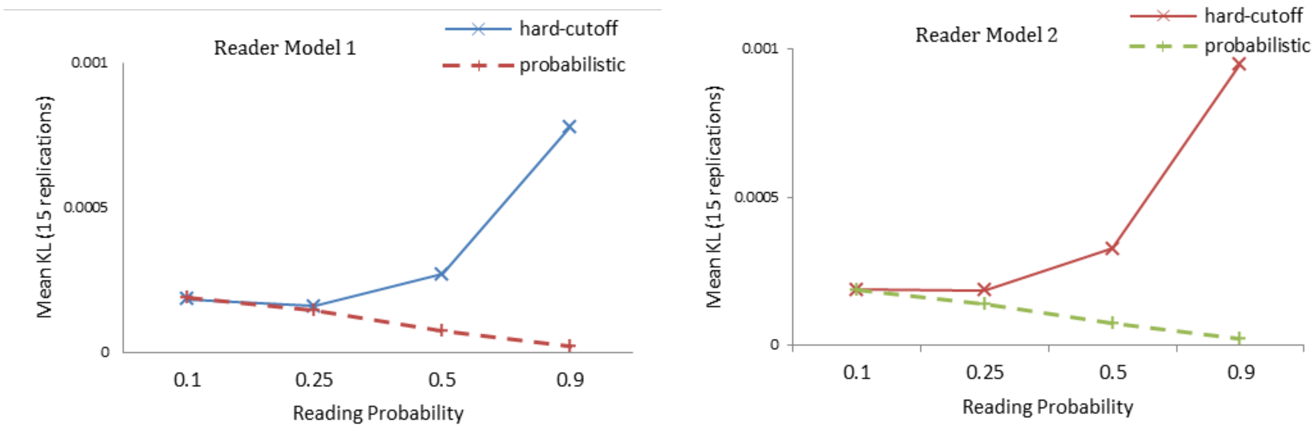

**Figure 2: Mean KL Distortion vs. Reading Probability (both Reader Models)**

Considering the performance based on MAE we observe that, Top-N seems to perform better than probabilistic NRS (figure 1), as the findings are established from both reader models in the simulation. However, under the second metric (KL) clearly probabilistic NRS outperforms Top-N NRS for both reader models (figure 2). These findings present the tradeoff between the two NRS in terms of an accuracy-distortion tradeoff. While probabilistic NRS seems to have a small accuracy loss (in terms of counts of articles it recommends) it is more true to the natural shares of the articles and does not create distortions which otherwise can occur.

### Comparison to an "Adapted" Influence Limiter Heuristic

Prior research has shown the benefits of probabilistic NRS when there is manipulation. However, the influence limiter algorithm (Resnick et al., 2007) is one that had been proposed in the past as a solution for some kinds of manipulation. This algorithm generates item recommendations controlling rater's influence on recommender systems through reputation acquired over time. The reputation of a rater is updated based on rating provided by him to an item and the $loss\ function$ determined through the prediction made to a target user compared to the actual preference of the target user.

In this research our focus has been on the counts of articles, and the reader's individual behavior (or reading pattern) has been left out for ease of exposition. Hence, the approach of Resnick et al. (Resnick et al., 2007) cannot directly be used. Instead, here we present a novel adaptation of this in our context to limit the influence of fake counts to generate article recommendations. In a similar vein it should be also noted that in the present analysis counts of articles is updated, instead of rater's reputation. We assume that the average time interval of two consecutive counts received by a recommended article is less than the average time interval of two consecutive counts received by the other articles in the system. A measure $\beta_{ij}$ has been introduced that limits the influence of a manipulator in the top$-N$ NRS. For any article $j$ at time $t_i$ it is defined as,

$$\beta_{ij} = \min(1, R_{ij}) = \min\left(1, \frac{t_i - t_0}{\alpha * (c_{ij} - 1)}\right) \qquad (1)$$

Influence limiting process operates between a pre-selected time intervals $(t_0, t_n)$; this can be determined through designer's experience or other appropriate choice can be the time interval when manipulation activity is most observed. For every $t_0 \le t_i \le t_n$ an article $j's$ reputation is updated as given in equation 1. In the expression $\alpha$ represents the average time interval that is "reasonable" between two consecutive counts received by a recommended article in the top$-N$ NRS (this can be determined through the arrival distribution of counts of the recommended articles) and, $c_{ij}$ is the number of counts received by the article $j$ in the time interval given by $(t_0, t_i)$. After $t_0$, at any given time point $t_i$ the new count received by the article $j$ (denoted as $c_{ij}'$) passes through an influence limiting process to generate a modified count given by $\tilde{c}_{ij}$ as described below in the pseudo code. ($\tilde{c}_{0j}$ represents count received before $t_0$ ). After $t_n$ each new count received by any of the articles is modified through its reputation $\beta_{nj}$ at time $t_n$. When $R_{ij} \ge 1$, all weight is on $c_{ij}'$ i.e. article $j$ has full credibility.

*An Adapted Influence Limiter Heuristic*

```
1. Get $\tilde{c}_{0j}$ for each article at $t = t_0$
2. For each article $j$, $c_{0j} \leftarrow 0$
3. For each $t_i$, when $1 \le t_i \le t_n$ and $c_{ij} \ge 2$
    a. For each article $j$
        i. $\beta_{ij} \leftarrow \min(1, R_{ij})$
        ii. $\tilde{c}_{ij} \leftarrow \tilde{c}_{i-1j} + \beta_{ij} * c_{ij}'$
        iii. $c_{ij} \leftarrow c_{i-1j} + c_{ij}'$
    b. End for
4. End for
```

Initially each article in the recommended list starts with initial reputation score of $R_{0j} = 1$. The current reputation $\beta_{ij}$ of an article limits the influence of excessive counts that the article has received on a given interval. The proposed algorithm discourages manipulators to extravagantly increase the count of target article. The reputation score $\beta_{ij}$ is always positive and bounded above by 1.

Let us consider the first user model in our simulation (when reader performs random selection of an article from the recommended list). Also it should be noted that in context of manipulation we are concerned about articles appearing in the recommended list. The initial 100 time steps have been selected as the observation period before implementing the modified count (the influence limiting heuristic) for each article. The selection of an article from the recommended list is performed randomly, hence the expected count that an article will receive over initial 100 time steps will be$\left(\frac{100}{10}\right) * p = 10 * p$, where $p$ the selected reading probability in the simulation. Hence, the expected time interval between two consecutive counts received by an article in Top$-N$ NRS is given by $\alpha = \left(\frac{100}{10*p}\right) = \frac{10}{p}$. Based on our choice of time period for the observation $t_i - t_0$ will be $t_i$. Hence, the reputation of an article $j$ at time $t_i$ will be given by (equation (1))

$$\beta_{ij} = \min(1, \frac{t_i * p}{10 * (c_{ij} - 1)})$$

As mentioned earlier, the major issue of interest is manipulation at the early stage (which will be very effective for the manipulator). Hence, variants of manipulation examined are heavy (50 clicks) and low (10 clicks) early manipulation. The articles of interest in this are also the $N^{th}$ and $(N+1)^{th}$ articles in the list.

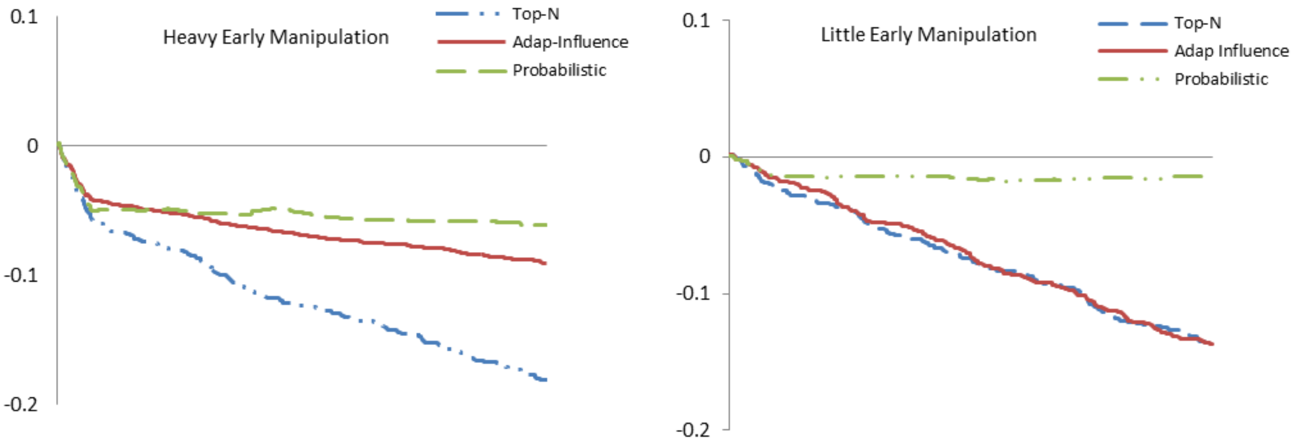

**Figure 3: Comparison of Manipulation based on M1**

The results based on the measure M1 (plotted on Y-axis) suggest that in the case of extreme manipulations, the proposed adapted influence limiter heuristic performs similar to probabilistic NRS (figure 3, left panel). This seems to be by the design of the adapted influence limiter heuristic - as the manipulator injects more fake counts for the target article, this leads to less reputation for it ($\beta_{ij}$). In turn new counts received by the manipulated article cause less cumulative increase in its count. However, small manipulation effort (especially if an article has just missed the cutoff for Top-N and the manipulator is in a position to determine this) may go undetected in case of the adapted influence limiter (figure 3, right panel). Here, probabilistic NRS is still robust.

## CONCLUSION

There has been growing evidence about the influence of web portals on news consumption behavior of readers. Due to participatory nature of news recommender systems, it is also giving rise to few active groups of users (Warren and Jurgensen, 2007; Lerman, 2007). This phenomenon is further propelled by penetration of social media, as the most popular stories cause a flood of traffic as recommendation ripple across social networks (The Economist, 2011). There are also some marketing companies also known to be in existence who promise customers to get the front page appearance in exchange of fee (Warren and Jurgensen, 2007). In other cases, there are companies known to be in existence who sell “tweets” and Facebook likes to gain popularity (e.g. pay4tweet.com, getfansfast.com).

In light of all this NRS should be careful to avoid common manipulation strategies, and also it should be able to generate diverse recommendation of articles. To address these issues in the present research we have proposed a simple probabilistic variant of count based NRS. The performance of the common Top-N news recommender and its probabilistic counterpart has been analyzed, based on two different metrics. Finally, an adapted influence limiter algorithm has been introduced, and its performance has been compared with probabilistic NRS. We show that the probabilistic NRS has practical implications in terms of providing a better way of utilizing information generated through users in comparison to the current Top-N NRS in the recommendation process. To our knowledge findings presented in this paper are unique contribution in the news recommendation systems research.

## REFERENCES

1. Chen, C., Wu, K., Srinivasan, V., and Zhang, X. (2011) Battling the Internet Water Army: Detection of Hidden Paid Posters. *ArXiv.org*, (November 18th).
2. Cho, J., Roy, S., and Adams, R. E. (2005) Page Quality: In Search of an Unbiased Web Ranking. *SIGMOD 2005*.
3. Dellarocas, C. (2006) Strategic Manipulation of Internet Opinion Forums: Implication for Consumers and Firms, *Management Science*, volume 52(10): 1577-1593.
4. Kullback, S., and Leibler, R. A. (1951) On Information and Sufficiency, *The Annals of Mathematical Statistics*, volume 22(1): 79-86.
5. Largillier, T., Peyronnet, G., and Peyronnet, S. (2010) SpotRank: A robust Voting System for Social News Websites. In *Proceedings of the 4th workshop on Information Credibility 2010*, pages 59-66.
6. Lee, J. S., and Zhu, D. (2011) Shilling Attack Detection-A New Approach for a Trustworthy Recommender System. *Informs Journal on Computing*, preprint March 10.
7. Lerman, K. (2007) Social Information Processing in News Aggregation. *IEEE Internet Computing*, volume 11(6): 16-28.
8. Lerman, K., (2007). User participation in social media: Digg study. In Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pages 255–258.
9. Resnick, P., and Sami, R. (2007) The Influence Limiter: Provably Manipulation-Resistant Recommender Systems, *Proceedings of ACM Conference on Recommender Systems (RecSys'07).*
10. Resnick, P., and Sami, R. (2008) The Information Cost of Manipulation-Resistance in Recommender Systems, *Proceedings of ACM Conference on Recommender Systems (RecSys'08).*
11. The Economist. (2011) Bulletins from the future: Special Report on the News Industry. July 7th 2011.
12. Warren, J., and Jurgensen, J. (2007). The Wizards of Buzz. *Wall Street Journal*, February 10.
13. Ziegler, C. N., McNee, S. M., Konstan, J. A., and Lausen G. (2005) Improving Recommendation Lists through Topic Diversification *(WWW 2005)*.
14. Prawesh, S. and Padmanabhan, B. (2011). The "top N" news recommender: count distortion and manipulation resistance. In *Proceedings of the fifth ACM conference on Recommender systems* (RecSys '11). ACM, New York, NY, USA, 237-244.