Measuring Coverage and Divergence of Reading Behaviors Among Friends*

Long T. Le
Rutgers University
longtle@cs.rutgers.edu

Tina Eliassi-Rad
Rutgers University
eliassi@cs.rutgers.edu

ABSTRACT

Given data from an online friendship network (e.g., the Facebook social network) and its social reader (i.e., a reading application deployed on a social network), how can we effectively capture the similarities between the reading behaviors of a user and her friends over time? Also, how can we effectively summarize such similarities across users? For the first question, we are interested in measures of *coverage* and *divergence*, which are based on tie-strength functions. Coverage captures the amount by which the first-order Markov assumption holds between the reading behaviors of user u and her friends. Divergence captures the amount of inconsistency in their *reading* tie-strength across time. We define *reading tie-strength* on a bipartite graph of $users \times articles$, where an edge $e(u, a)$ indicates that user u read article a. In this work, we study three popular tie-strength functions (namely, Common Neighbor, Jaccard Index, and Adamic-Adar); and propose coverage and divergence measures for the social-reader domain. In addition, we introduce a method for summarizing coverage and divergence values across all users that takes into account the heavy-tailed distribution and the sparsity of the data. Our experiments on a real-world dataset from a large media company demonstrate that Common Neighbor (i.e., number of common articles read between a user and her friend) is a better tie-strength function for the social-reader domain than Jaccard Index or Adamic-Adar.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; E.1 [**Data Structures**]: Graphs and networks

*We thank Matthew Bryan and Robert Chang for providing the data for this work and their useful comments. This work was supported in part by Washington Post Labs, NSF CNS-1314603, DTRA HDTRA1-10-1-0120, and DAPRA under SMISC Program Agreement No. W911NF-12-C-0028.

NewsKDD '14 New York, NY USA

General Terms

Design, Measurement, Performance, Experimentation

Keywords

Social reading, tie strength, social ranking

1. INTRODUCTION

We address two problems w.r.t. social news reading applications. For both problems, we are given (*i*) an online friendship network (such as the Facebook social network) and (*ii*) the data from a social-reader application deployed on that friendship network. Popular social readers include the Washington Post Social Reader and the Guardian Social Reader. In the first problem, we are interested in quantifying similarities between the reading habits of a user and her friends over time. In the second problem, we are interested in methods that effectively summarize such similarities across users. The motivation behind these problems are to better understand the activities on a social reader. This newly gained understanding can then be used to devise better algorithms that promote application engagement.

Problem 1: Quantifying similarities between the reading habits of a user and her friends over time. We present two measures: *coverage* and *divergence*. Both measures are based on tie-strength functions [6]. Coverage captures how much the first-order Markov assumption holds between the reading behaviors of user u and her friends—i.e., the normalized *reading* tie-strength between user u and her friends at time t_{i+1} assuming that they had a positive *reading* tie-strength at time t_i. Divergence captures how much inconsistency exists in their *reading* tie-strength over time. We define *reading* tie-strength as the value that a tie-strength function outputs when given a bipartite graph of $users \times articles$. An edge in this bipartite graph represents an article read by a user. Our coverage and divergence measures are *local*; thus to compute them, we restrict the set of nodes in the bipartite graph (of $users \times articles$) to just the user of interest, her friends, and (the union of) the articles they read. Tie-strength functions that have high coverage and low divergence are considered more effective than ones that have low coverage and high divergence. W.r.t. particular tie-strength functions, we investigate *Common Neighbor*, *Jaccard Index*, and *Adamic-Adar*. Section 2 contains formal definitions of coverage and divergence.

Problem 2: Summarizing similarities across users. The solution to Problem 1 gives us two values per user: one for coverage and another for divergence. We need a method

for aggregating these values across a set of users. A naive way of summarizing the coverage and divergence values is to average them across all users. This naive method has two limitations. It does *not* address (*i*) the fact that the degree distribution (over friendships) is power-law with a heavy tail; and (*ii*) the prominent data-sparsity issue whereby many users have zero reading tie-strength with their friends. In Section 2, we propose a method that addresses these limitations by creating *aggregate* data structures, and by taking advantage of the inherent taxonomy in articles (e.g., an article about football falls under sports).

Our **contributions** are threefold: **(1)** We introduce coverage and divergence measures, which are based on reading tie-strengths among users, to quantify similarities between the reading habits of users. **(2)** We introduce a method that effectively summarizes the coverage and divergence measures across users. **(3)** Our extensive empirical study on real-world data from a large media company demonstrates that some tie-strength functions (such as Common Neighbor) are better suited for social news reading applications that others (such as Jaccard Index or Adamic-Adar); and that operating at the topic-level (such as sports) is more effective than the article-level.

This paper is organized as follows: Section 2 (Proposed Method), Section 3 (Experiments), Section 4 (Related Works), and Section 5 (Conclusions and Future Work).

2. PROPOSED METHOD

This section is divided into three parts: preliminaries, our measures of coverage and divergence, and our summarization method.

2.1 Preliminaries

Articles have meta data. For each article, we know what is its editor-assigned *topic* (e.g., football) and what is its editor-assigned *section* (e.g., sports). In our real-world data (see Section 3), we have 6,024 topics and 11 sections. Each article is assigned to *at least* one topic; a topic is assigned to *only one* section; and a section has *many* topics.

As discussed in Section 1, our measures of coverage and divergence operate on bipartite graphs of $user \times articles$, where an edge $e(u, a)$ indicates that user u read article a. Since we are interested in quantifying similarities between the reading behaviors of a particular user u and her friends, our bipartite graphs are *local*. This means that the nodes in our bipartite graph are u, her friends, and the collection of articles read by them. Some of these bipartite graphs are extremely sparse (with very few edges). In such cases, we somewhat alleviate the sparsity of our graphs by restricting the article nodes to articles on a particular topic. Alternatively, we replace the article nodes by topic nodes; and have bipartite graphs $users \times topics$, where an edge $e(u, c)$ indicates that user u read *any* article on topic c. Thus, for our real-world data with 11 sections, we generate 24 bipartite graphs for each user u and each time period:

- G_1: (user u & friends) $\times$ (articles read across all sections)
- G_2-G_{12}: (user u & friends) $\times$ (articles read in section s)
- G_{13}: (user u & friends) $\times$ (topics read across all sections)
- G_{14}-G_{24}: (user u & friends) $\times$ (topics read in section s)

As we show in Section 3, the choice of the bipartite graph affects the coverage and divergence values. For instance, operating at the topic-level (i.e., G_{13} through G_{24}) consistently produces better performance (in terms of higher coverage and lower divergence values) than operating at the article-level (i.e., G_1 through G_{12}). This is because the bipartite graphs at the topic-level are less sparse than those at the article-level.

2.2 Coverage and Divergence

Given a time period t_i, we construct the aforementioned bipartite graphs for each user u. We then give these bipartite graphs to tie-strength functions to compute the *reading* tie-strength TS between user u and her friends. In this work, we study three tie-strength functions−namely, CN: Common Neighbor; JI: Jaccard Index; and AA: Adamic-Adar [1]. We selected these three TS functions based on their popularity and our previous work in [6].

For a node u, we use $\Gamma(u)$ to denote the set of articles that u read. For an article P, we use $|P|$ to denote the number of people that read P. The formal definitions of tie-strength measures used in our study are as follows. **Common Neighbor (CN)**: The tie strength between u and v is equal to the total number of articles that both u and v read: $TS_{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|$. **Jaccard Index ($JI$)**: This tie-strength measure is a normalized version of CN, where we normalize for the reading tendencies of u and v: $TS_{JI}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$. **Adamic-Adar ($AA$)**: This tie-strength measure increases as users u and v read more common articles; and it discounts for very popular articles: $TS_{AA}(u, v) = \sum_{P \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |P|}$.

Thus far, for all users u and for each time period t_i, we have built a set of bipartite graphs and have calculated the tie strength between user u and her friends. Next, we present the formal definitions of coverage and divergence. For this discussion, let $TS(u, k, t_i)$ denote the tie strength between u and her k^{th} friend during time period t_i.

Coverage computes the normalized tie-strength between user u and her friends at time t_{i+1} assuming that they had a positive tie-strength at time t_i (see Equation 1). If a tie-strength function generates high coverage values, then it is considered an effective function for capturing reading habits in a social reader over time. We formally define coverage as follows:

$$Cov(u, t_i, t_{i+1}) = \frac{\sum_{k=1}^{n_u} t_i[u, k] \times TS(u, k, t_{i+1})}{\sum_{k=1}^{n_u} TS(u, k, t_{i+1})} \quad (1)$$

where n_u = number of u's friends; and $t_i[u, k] = 1$ when $TS(u, k, t_i) > 0$. Our coverage formula captures the magnitude of tie strength between u and her k^{th} friend at time t_{i+1} given that they had a positive tie-strength at time t_i. We also tried other coverage formulas such as capturing the minimum tie-strength across t_i and t_{i+1}. They underperformed compared to the coverage formula in Equation 1. For brevity, we have omitted them.

Divergence computes the amount of inconsistency in the reading tie-strength between time periods by using the *normalized Canberra distance* (see Equation 2). Canberra distance is sensitive to small changes near zero, and it normalizes the absolute difference of the individual comparisons. If a tie-strength function generates low divergence values, then it is considered an effective function for tracking the reading habits in a social reader.

$$Div(u, t_i, t_{i+1}) = \frac{1}{n_u} \sum_{k=1}^{n_u} \frac{|TS(u,k,t_i) - TS(u,k,t_{i+1})|}{|TS(u,k,t_i)| + |TS(u,k,t_{i+1})|} \quad (2)$$

There is an inherent trade-off between coverage and divergence. The higher the coverage, the lower the divergence; and vice versa. We use the *harmonic mean H* of coverage and one minus divergence to summarize coverage and divergence into one measure (see Equation 3).

$$H(Cov, (1 - Div)) = \frac{2 \times Cov \times (1 - Div)}{Cov + (1 - Div)} \quad (3)$$

2.3 Method for Summarizing Coverage and Divergence Across Users

Now that we have coverage and divergence values for all users, we need a method for summarizing them across users. As discussed earlier, merely averaging across the users is not a good approach since our data is heavy-tailed and sparse. A better way of summarizing coverage and divergence is to compute them for a *super user* instead of individual users. Our procedure is as follows. For each bipartite graph representation (see Section 2.1) and for each user u, we have a tie-strength table whose rows are the friends of u and whose columns are the tie strengths between u and her friends for various time periods. We create a tie-strength table for a *super user* by concatenating the tie-strength tables of all users. The rows in the super user's tie-strength table are *friendship pairs*. The columns are the tie strengths of each friendship pair for different time periods. The super user's tie-strength table is less sparse than any individual tie-strength table because the number of rows in the former is much larger than the latter, while the number of columns is the same. Additionally, by computing the coverage and divergence values on the super user's tie-strength table, we differentiate between more popular and less popular users, whereby the more popular users get more weight in coverage and divergence computations.

Like people, some sections are more popular than others. For example, Arts & Entertainment is more popular than Science. We use the same procedure as above to address this issue. In particular, we concatenate the super-user tie-strength tables from various sections and create a *great* table. The rows in this great table are *friendship pairs* for a given section (i.e., a triple $\langle u$, friend of u, section $s\rangle$) and the columns are the tie strengths of each triple for various time periods. This great table is less sparse than each section's super-user tie-strength table. Also, the computation of convergence and divergence on the great table's tie-strength values gives more weight to more popular sections.

As discussed earlier, our bipartite graphs are sparse (especially the ones whose article nodes represent articles-read in a particular section). To alleviate this sparsity, we utilize the bipartite graphs whose article nodes are articles-read across all sections. Specifically, if we do not have observations (i.e., article-reads) to compute coverage and divergence based on an article in a particular section, we grab the observations from the articles across all sections. The computation here will be at a coarser level and will produce a higher coverage at the expense of lower divergence values; but they allow us to compute coverage and divergence for everyone.

3. EXPERIMENTS

Data Description. Our data (from a large media company) contains three sets of information: (1) an anonymized social network with 37M users and 502M friendships; (2) a reading trace containing 104M records that capture reading activity across seven months; (3) category designations by editors, who categorize articles into 11 different sections: Arts & Entertainment, Business, Education, Family & Society, Health, Life & Style, News, Recreation, Science, Sports, and Technology. Each section is divided into a set of topics. There are a total of 6,024 topics. Similar articles are grouped into a topic; and similar topics are grouped into a section. As expected, the degree distribution of the social network is power-law with a heavy tail (e.g., one user has approximately 5K friends). Also, as expected, the distribution over the number of articles-read is power-law with a heavy tail (e.g., one user read 18.1K articles over the span of 7 months). For brevity, we omit the plots for these distributions. Figure 1 depicts the distribution of topics, articles, and readings in each section. The News section has the highest number of topics and articles; the Arts & Entertainment section has the highest number of readings.

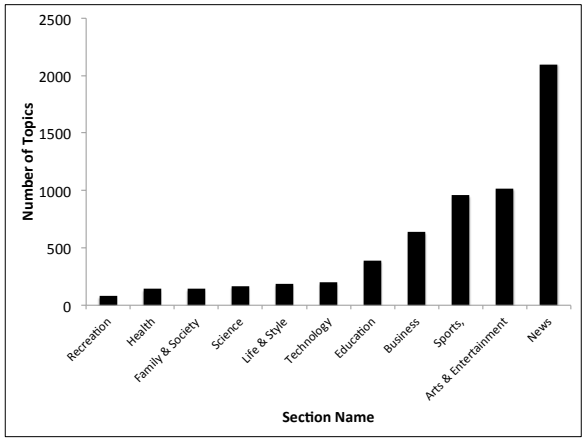
Like all real-world data, our data is noisy. For brevity, we omit the description of our data cleaning. Our results are based on a set of 1,829 most active readers, who have at least 20 friends. We divide the timeline into 6 periods, with equal number of articles read in each time period.

Results. For brevity, we only show two of our results. Figure 2(a) depicts the summary of various harmonic means of coverage and 1−divergence, with different bipartite-graph representations. The topic level (i.e., bipartite graphs of $users \times topics$, where an edge represents a user reading *any* article in a topic) produces better performance in terms of (higher) harmonic means than the article level (i.e., bipartite graphs of $users \times articles$, where an edge represents a user reading a particular article). Common Neighbor (CN) has the highest harmonic mean in both the article- and topic-level. Figures 2(b) and (c) plot coverage and divergence of Common Neighbor across time periods, with different bipartite-graph representations. Similar to Figure 2(a), we observe that bipartite graphs at the topic level (such as topics in a particular section) have higher coverage and lower divergence than the bipartite graphs at the article level.

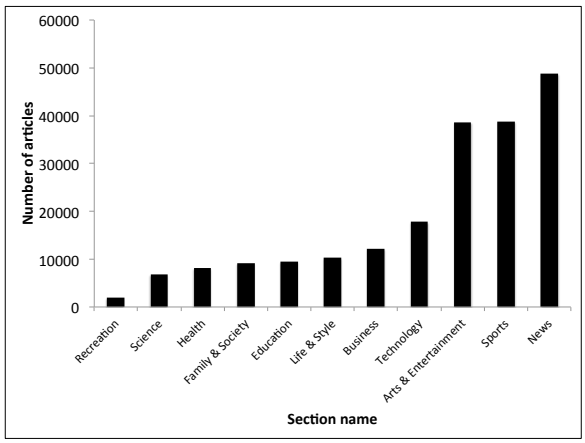
Discussion. Common Neighbor has better performance, in terms of coverage and divergence, than Jaccard Index and Adamic-Adar. Common Neighbor is a computationally efficient tie-strength function; and it is intuitive. Representing social reading activity at the topic level (i.e., $users \times topics$) yields better results than at the article level (i.e., $users \times articles$). Capturing divergence is a harder task than capturing coverage.

4. RELATED WORK

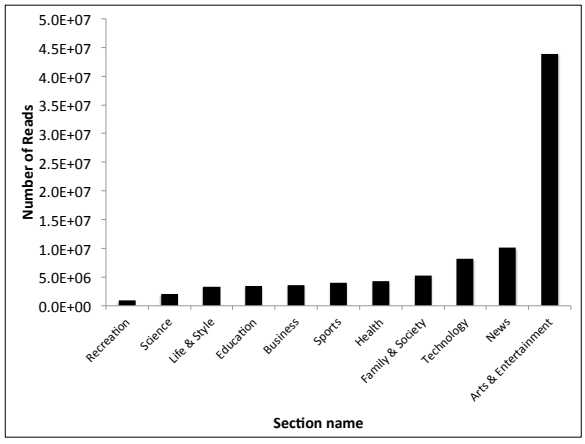
To the best of our knowledge, no one has used tie-strength functions to define coverage and divergence measures in a social reader for the task of quantifying similarities between users' reading habits. Some related literature include work on social influence in various online advertising domains [8, 2, 9, 3]. In social readers, we expect social influence to play a big role in whether a user reads an article. Of course, other factors such as how prolific of a reader the user is or how popular the article is also play important roles.



(a) Distribution of Topics in Each Section

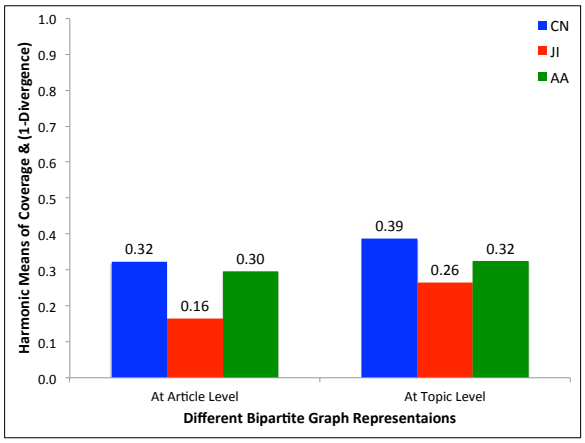


(b) Distribution of Articles in Each Section

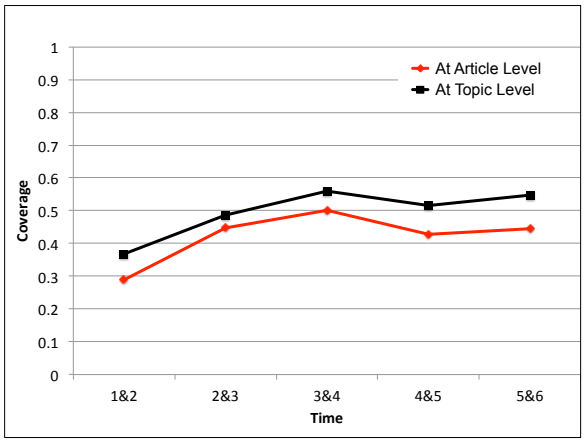


(c) Distribution of Readings in Each Section

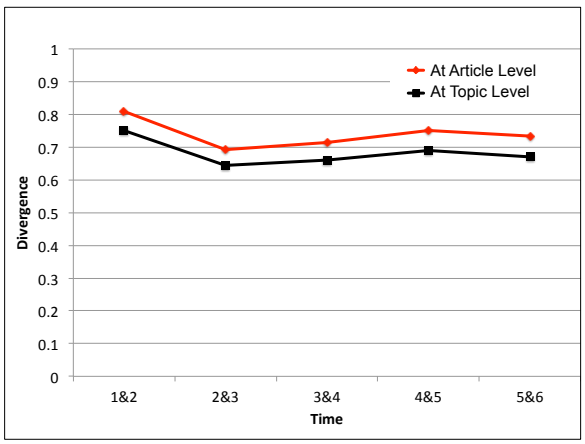
Figure 1: Distributions of topics/articles/readings in each section. (a) and (b) show that News has the highest number of topics and articles; but (c) shows that most of articles read (on the social reader) are in Arts & Entertainment.



(a) Harmonic Mean of Coverage and 1−Divergence.



(b) Coverage of Common Neighbor Over Time



(c) Divergence of Common Neighbor Over Time

Figure 2: (Best viewed in color) Various performance measures with different bipartite graph representations. Bipartite graphs at the topic level (i.e., *users* × *topics*) yield better performance in terms of both coverage and divergence than those at the article level (i.e., *users* × *articles*). CN has higher harmonic means than JI and AA.

The research community has long been interested in measuring the strength of ties between people. For instance, Granovetter [5] discussed the strength of weak ties, where acquaintances are more likely to be "strong bridges" than best friends. Gupte and Eliassi-Rad [6] introduced an axiomatic approach to infer implicit weighted social networks from bipartite graphs. Additionally, tie strength is closely related to link prediction [1, 10] and prediction of the strength of ties in social networks [7, 4].

5. CONCLUSIONS AND FUTURE WORK

We introduced two problems w.r.t. social readers: (1) quantifying similarities between the reading behaviors of a user and her friends; and (2) summarizing these similarities across users. For the former problem, we presented coverage and divergence measures based on the reading tie-strengths among users. For the latter problem, we described a method that effectively aggregates coverage and divergence values across users. Our experiments on real-world data from a large US-based media company demonstrated that it is more effective to operate at the topic-level than at the article-level; and showcased that some tie-strength functions (such as Common Neighbor) are better suited for social news reading applications than others (such as Jaccard Index or Adamic-Adar).

Future work. Users often have many "friends" on online social networks (since the burden of friendship is low), it would be better to only track their top K friends per topic. Also, clustering friends into similar groups has the potential to improve performance in terms of coverage and divergence. Lastly, integrating various sources of information to alleviate sparsity could improve results.

6. REFERENCES

[1] L. A. Adamic. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2003.

[2] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence*, pages 207–214, 2005.

[3] E. Bakshy, D. Eckles, R. Yan, and I. Rosenn. Social influence in social advertising: Evidence from field experiments. In *EC*, pages 146–161, 2012.

[4] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *CHI*, pages 211–220, 2009.

[5] M. Granovetter. The strength of weak ties. *American J. of Sociology*, 78(6):1360–1380, 1973.

[6] M. Gupte and T. Eliassi-Rad. Measuring tie strength in implicit social networks. In *WebSci*, pages 109–118, 2012.

[7] I. Kahanda and J. Neville. Using transactional information to predict link strength in online social networks. In *ICWSM*, 2009.

[8] P. Monge and N. Contractor. *Theories of Communication Networks*. OUP, 2003.

[9] P. Papadimitriou, H. Garcia-Molina, P. Krishnamurthy, R. A. Lewis, and D. H. Reiley. Display advertising impact: Search lift and social influence. In *KDD*, pages 1019–1027, 2011.

[10] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, and R. Merom. Suggesting friends using the implicit social graph. In *KDD*, pages 233–242, 2010.