Topical Organization of User Comments and Application to Content Recommendation

Vidit Jain
Yahoo! Labs Bangalore, India
viditj@yahoo-inc.com

Esther Galbrun
Department of Computer Science and
Helsinki Institute for Information Technology HIIT,
University of Helsinki, Finland
galbrun@cs.helsinki.fi

ABSTRACT

On a news website, an article may receive thousands of comments from its readers on a variety of topics. The usual display of these comments in a ranked list, e.g. by popularity, does not allow the user to follow discussions on a particular topic. Organizing them by semantic topics enables the user not only to selectively browse comments on a topic, but also to discover other significant topics of discussion in comments. This topical organization further allows to explicitly capture the immediate interests of the user even when she is not logged in. Here we use this information to recommend content that is relevant in the context of the comments being read by the user. We present an algorithm for building such a topical organization in a practical setting and study different recommendation schemes. In a pilot study, we observe these comments-to-article recommendations to be preferred over the standard article-to-article recommendations.

Categories and Subject Descriptors

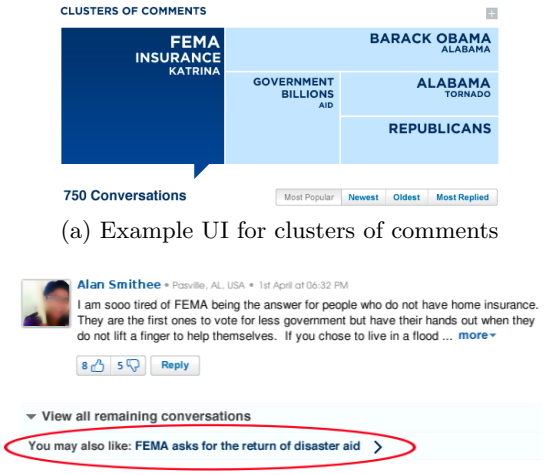
H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Algorithm; User Generated Content; Recommendation;

1. INTRODUCTION

Consider the example of an article about a recent tornado in Alabama. The main topics appearing in the comments for this article are displayed in a compact and interactive interface (see Figure 1(a)). Each block corresponds to one cluster, summarized by a tag cloud of the entities that occur in it. In this illustration, the user is reading the comments that discuss the emergency responses of the Federal Emergency Management Agency (FEMA) after this natural disaster (cluster highlighted in a darker shade). At a different time, this user may be reading other comments that question the legitimacy of government's spending of billions of dollars as foreign aid. In the first case, the user is likely to be more interested in reading about the responses after similar natural disasters, such as hurricane Katrina, as opposed to information on foreign aid efforts. Existing systems either use only the article text along with a global personalization model [1], or enrich the parent article with a holistic analysis of all of the associated comments [2] to recommend articles. These systems fail to distinguish between the above two different information needs of a single user. Our system addresses this issue by recommending an article relevant to the specific context of a semantically coherent cluster of comments (Figure 1(b)).



(a) Example UI for clusters of comments

(b) Recommendation is shown below the comments

Figure 1: ***Recommending content in the context of user comments.***

Our conclusions do not compete with, but complement, other approaches that model user profiles [1] and user similarities [3] for providing personalized recommendations.

2. TOPICAL ORGANIZATION

Building topical organization for user comments has its unique challenges. The difference in the language skills and seriousness of comment writers leads to a huge variation in the language and content across comments. The standard natural language processing and information extraction systems perform poorly on these noisy, short pieces of text. Also, for popular articles, the comments arrive at a high rate and the topics of discussions evolve rapidly over time. On a typical news website, the desired system need to compute these topics for thousands of articles every few minutes. Another constraint is due to the actual representation of the extracted topics to the users – in order to provide a consistent user experience across different articles, the hosting website prescribed that at most *five* different clusters

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

	TE	TO	CE	COV
k-means	0.542	**4.353**	0.120	0.474
METIS	0.517	23.003	0.858	**0.511**
EBC	**0.511**	14.960	**1.912**	0.505

Table 1: *Comparison of clustering algorithms.* Average values of the entropy of intra-cluster term frequencies (TE), per article inter-cluster overlap between top-5 terms (TO), entropy of topic sizes (CE), and fraction of comments assigned a topic (COV).

be shown per article. While it is interesting to study other choices of number of topics as well, we focus only on obtaining up to five topics per article in this work. In addition, the usual properties (i.e., intra-topic semantic coherence, inter-topic semantic distinction, and balanced distribution of topic sizes) for any topical organization are desired.

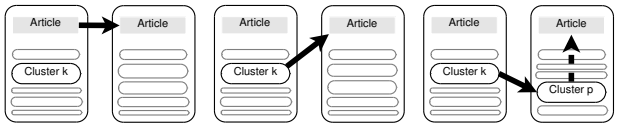
Here we compute topics as clusters of comments. Even though individual comments show diverse and unreliable language, the entities mentioned in them provide a useful signal for clustering comments to form topics. Here they are identified using a system [4] optimized for news domain. The set of these entities, however, varies significantly across articles and the comment count for different entities follows a heavy-tailed distribution. Hence a global set of important entities is ineffective for clustering comments into topics across articles; the topics need to be computed separately for each article from *all* of its corresponding comments.

Entity-based clustering (EBC). To address the latency and other requirements mentioned above, we designed an algorithm that highlights entities. We first extract the named-entities and selective parts of speech, e.g. nouns and adjectives from the comments. While clustering, only these extracted terms are used to represent the comments. Then, we partition the comments into small initial subsets based solely on the occurrences of named-entities. Next, we construct a graph where each vertex represents an initial subset. The weight of the edge between every pair of vertices is computed as the inner product of the two tf-idf representations of the comments belonging to the corresponding partitions. Finally, the METIS k-way partitioning algorithm[1] is used to obtain *five* partitions of this graph as the clusters of comments. Table 1 shows a comparison of EBC with other computationally feasible alternatives on 19,320 articles, each of which received at least 100 comments. Based on these observations, we choose EBC for the rest of our experiments.

3. CONTENT RECOMMENDATION

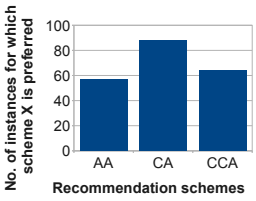
We consider three recommendation schemes (see Figure 2). The traditional setting for article recommendation (AA) selects the best matching article for a given source article. The second scheme (CA) selects the best matching article for the comments belonging to a given source cluster of comments. Finally, the third scheme (CCA) finds the best matching cluster of comments and selects the parent article of this matching cluster as the target recommendation. These schemes are implemented using a retrieval engine similar to Lucene, which allows a rapid indexing of large volumes of textual data. We use the standard tf-idf representation for the news articles. However, since the comments in the topical clusters show unusual repetitions of terms, we use a slightly different tf term for clusters of comments, i.e., $\text{tf}_{clu}(t,c) = \frac{\#\text{ comments in cluster } c \text{ containing term } t}{\#\text{ comments in cluster } c}$. The inverse document frequencies of different terms are computed separately for articles and clusters, and denoted by $\text{idf}_{clu}(t)$ and $\text{idf}_{art}(t)$. The relevance of an indexed document d to a source document s is then defined as $R(s,d) = \sum_{t \in d} \text{tf}_T(t,s) * \text{tf}_{RS}(t,d) * \text{idf}_T(t)^2$, where $T \in \{clu, art\}$ is the type of the source document and tf_{RS} the term frequency computed by the retrieval engine.

[1] www.cs.umn.edu/∼metis



(a) AA (b) CA (c) CCA

Figure 2: *Different recommendation schemes.*



Accepted hypothesis	p-value
CA ≻ AA	0.003
CA ≻ CCA	0.031

Figure 3: *Observations from the pilot study.* P-values computed for 1-sided χ^2 statistical significance test.

We collected the comments for 1884 popular articles (each with at least 500 user comments) from Yahoo! News. In all, we obtained about 9000 clusters (over 2 million comments), for each of which we generated the recommendations using the above schemes. In our pilot study, we first asked users to read comments from a (source) topic. Then they were shown recommendations from each of the above schemes, and asked to choose the most relevant one in the context of the source cluster. As seen in Figure 3, CA is preferred significantly more than both of the AA and CCA schemes. We observed that using topical clusters as sources enables us to identify engaging topics that are latent in the article but are made explicit in the comments. We further investigated machine-learned ensemble techniques that select one of these three recommendations based on the source document. As shown in Table 2, a Gaussian-process based ensemble method was found to perform the best.

Multi-class SVM	Naïve-Bayes	GP-ensemble
-3.6%	+1.8%	**+8.5%**

Table 2: *Improvement (over* CA*) in predicting user preferences using ensemble methods.*

4. REFERENCES

[1] D. Agarwal et al. Personalized recommendation of user comments via factor models. In EMNLP, 2011.

[2] J. Wang et al. User comments for news recommendation in social media. In SIGIR, 2010.

[3] G. Linden et al. Amazon.com recommendations: item-to-item collaborative filtering. Internet Computing, 2003.

[4] A. McCallum et al. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In CONLL, 2003.