Interest Analysis using PageRank and Social Interaction Content

Chung-chi Huang
Institute of Information Science
Academia Sinica
Taipei, Taiwan 115
u901571@gmail.com

Lun-Wei Ku
Institute of Information Science
Academia Sinica
Taipei, Taiwan 115
lwku@iis.sinica.edu.tw

Abstract

We introduce a method for learning to predict reader interest. In our approach, social interaction content and both syntactic and semantic features of words are utilized. The proposed method involves estimating topical interest preferences and determining the informativity between articles and their social content. In interest prediction, we integrate articles' *quality* social feedback representing readers' opinions into articles to get information which may identify readers' interests. In addition, semantic aware PageRank is used to find reader interest with the help of word interestingness scores. Evaluations show that PageRank benefits from proposed features and interest preferences inferred across articles. Moreover, results conclude that social interaction content and the proposed selection process help to accurately cover more span of reader interest.

1 Introduction

Web keyword extraction tools such as KEA (www.nzdl.org/Kea/) typically look at articles from authors' perspective to calculate the importance of a word in articles. However, keywords are not necessarily words that interest readers. We found that articles could be analyzed more towards reader interest if a system exploited social interaction content (e.g., reader feedback) in social media.

Consider the content of an example article. The Web post describes a newly-renovated old house and the history, life style, and surrounding sightseeing sites of a historical city where it is located. Most keyword tools can easily identify keywords *the old house* (*謝宅*) and *the historical city* (*台南*). However, article readers might also be interested in less frequent words like *life style* (*生活*) and *traditional market* (*市場*), and single-occurrence like *rental fees* (*費用*), which are also mentioned in most reader feedback.

In the proposed method, an article was transformed into a word graph where vertices were words in the article and edges between vertices indicated words' co-occurrences. To distinguish associate/key words from words of reader interest, readers' *quality* interaction feedback was considered when building the word graph. Subsequently, word interest preferences and PageRank were utilized to find interest terms. Weightings concerning syntactic and semantic features are utilized in PageRank. Moreover, content-source and content-word weighted PageRank were exploited to return words for interest evaluation. The predicted interests can further be used as candidates for social tagging or article recommendation.

2 Related Work

The state-of-the-art keyword extraction methods have been applied to a myriad of natural language processing tasks including document categorization and summarization (Manning and Schutze, 2000; Litvak and Last, 2008), indexing (Li et al., 2004), information retrieval (Turney, 2000), and text mining on social networking or micro-blogging services (Li et al., 2010; Zhao et al., 2011;Wu et al., 2010). Here we extract keywords related to readers' interests.

Recently, collaborative tagging or social tagging has grown in popularity among Web services and received much attention (Golder and Huberman 2006; Halpin et al., 2007). Instead of analyzing user (tagging) activity or tag frequencies, we analyze articles and their social interaction content to predict reader interests.

Researches have been done on reader profiling for content recommendation. White et al. (2009) examined five types of contextual information in website recommendation while Ye *et al.* (2012) further explored social influence on item recommendation. Moreover, Tsagkias and Blanco (2012) concentrated on analyzing users' browsing behavior on news articles, and Jin (2012) recommended contents through a unified, per-

International Joint Conference on Natural Language Processing, pages 912–916,
Nagoya, Japan, 14-18 October 2013.

sonalized messaging system. In our work, the accumulated social interaction content is utilized to help determine the interest of future reader.

In studies more related to our work, Liu (2010) and Zhao (2011) present PageRank for keyword analyses using article topic information. The main difference from our current work is that we integrate social content and (global) topical interest preferences for words into (local) content-word weighted PageRank algorithm.

3 Finding Interests

To introduce the finding process of interests, we start from the problem statement. Given an article collection of various topics from social media (e.g., blogs), an article *ART*, and its reader feedback *FB*, our goal is to determine a set of interest words that are likely to represent the interest of future readers after reading *ART*.

3.1 Estimating Topical Interest Preferences

Basically, the estimation of topical interest preferences is to calculate the significance or degree of references of a word in a domain topic. The learning process contains four stages: (1) Generate article-word pairs in training data, (2) Generate topic-word pairs in training data, (3) Estimate interest preferences for words w.r.t. article topics based on different strategies, and (4) Output word-and-interest-preference-score pairs for various estimation strategies. In the first two stages of the learning process, we generate two sets of article and word information. The input to these stages is a set of articles with author-chosen topics and, if any, their reader feedback responses. The output is a set of pairs of article ID and word in the article, e.g., (*art*=1, *w*="old house"), and a set of pairs of article topic and word in the article, e.g., (*tp*="travel", *w*="old house"). Note that the article referred here may or may not contain the social reader feedback (See Section 4). In the third stage, we utilize aforementioned sets to estimate reader interest preferences for words across articles and across domain topics. Six different estimation strategies are as follows.

tfidf. The first estimation is a traditional yet powerful one, tfidf (term frequency multiplied by inversed document frequency):

$$\text{tfidf}(art, w) = \text{freq}(art, w) / \text{artFreq}(w).$$

Pr(w|tp). The second leverages a word's Maximum Likelihood Estimation under a given topic:

$$\Pr(w \mid tp) = \text{freq}(tp, w) / \sum_{w'} \text{freq}(tp, w').$$

Pr(tp|w). The third computes the topic-wise senses of a word:

$$\Pr(tp \mid w) = \text{freq}(tp, w) / \sum_{tp'} \text{freq}(tp', w).$$

entropy. The fourth is entropy which utilizes the uncertainty in topics to estimate its topic spectrum or its topic focus:

$$\text{entropy}(w) = -\sum_{tp'} \Pr(tp' \mid w) \times \lg(\Pr(tp' \mid w)).$$

Pr-Entropy(w|tp). The fifth further considers topic uncertainty in MLE: $\Pr(w \mid tp) / 2^{\text{entropy}(w)}$.

Pr-Entropy(tp|w). The last is a combination of the third and the fourth: $\Pr(tp \mid w) / 2^{\text{entropy}(w)}$.

These six estimations all take global information (i.e., article collection) into account.

3.2 Predicting Interest for Future Reader

Reader interests were predicted using the procedure in Figure 1. In this procedure we exploit semantic aware RageRank and reader feedback in social media to evaluate readers' interest in an article word. According to our observations, the collection of the reader feedback may reveal the common interest and browse habits of potential readers of the same article.

However, not all reader feedback responds to the article. Therefore, we screen reader feedbacks in Step (1) based on the article *ART*, its feedbacks *FB* and interest preference scores *IntPrefs*. The algorithm for identifying reader responses of a good quality, called *quality* reader responses hereafter, is as follows.

(1) *ngramsart=generateNgram(ART)*
(2) *Focused=findFocused(IntPrefs)*
(3) *selectedSt=NULL*
for each sentence st in FB
(4a)*ngramsst = generateNgram(st)*
(4b)*informativityco = Coverage-evaluate (ngramsst , ngramsart)*
(4c)*informativityfo = Focus-evaluate (ngramsst , Focused)*
(4d)*append st into selectedSt if conditions hold*
return selectedSt

Each response is evaluated at sentence level concerning informativity checked in two aspects. The first concerns the topic cohesion between reader response sentence *st* and article *ART*. Similar to BLEU's (Papineni et al., 2002) weighted ngram precision in machine translation, we compute the weighted ngram coverage of *st* (Step (4b)) on *ART* and favor the coverage of longer grams. Larger ngram coverage indicates higher topic correlation between the two. The second considers the topic distributions of words in *st*. We first rank and identify the words expected to have low topic uncertainty. Entropy estimation in

Section 3.2 is used for this purpose to find *Focused* (Step (2)). Then the informativity on topic focus of *st* is computed as the percentage of its words in set *Focused*. In the end, we prune reader sentences in *FB* according to the thresholds set for $informativity_{co}$ and $informativity_{fo}$ (Step (4d)).

After incorporating quality feedback *qualityFB* into *ART* (Step (2) in Figure 1), we construct a word graph for both the article and social content. The word graph is represented by a *v*-by-*v* matrix **EW** where *v* is the vocabulary size. **EW** stores normalized edge weights for word w_i and w_j (Step (4) and (5)). Note that the graph is directional from w_i to w_j and that edge weights are the words' co-occurrence counts within window size *WS*.

```
procedure PredictInterest(ART,FB,IntPrefs, λ , α ,N)
(1) qualityFB=selectInformativeFB(ART,FB,IntPrefs)
(2) Concatenate ART with qualityFB into Content
//Construct word graph for PageRank
(3) EW_{v×v}=0_{v×v}
    for each sentence st in Content
      for each word w_i in st
        for each word w_j in st where i<j and j-i ≤ WS
          if not IsContWord(w_i) and IsContWord(w_j)
(4a)        EW[i,j]+=1×m×srcWeight
          elif not IsContWord(w_i) and not IsContWord(w_j)
(4b)        EW[i,j]+=1×(1/m)×srcWeight
          elif IsContWord(w_i) and not IsContWord(w_j)
(4c)        EW[i,j]+=1×(1/m)×srcWeight
          elif IsContWord(w_i) and IsContWord(w_j)
(4d)        EW[i,j]+=1×m×srcWeight
(5) normalize each row of EW to sum to 1
//Iterate for PageRank
(6) set IP_{1×v} to
          [IntPrefs(w_1), IntPrefs(w_2), …,IntPrefs(w_v)]
(7) initialize IN_{1×v} to [1/v,1/ v, …,1/v]
    repeat
(8a)  IN'= λ × IN× EW+(1- λ )× IP
(8b)  normalize IN' to sum to 1
(8c)  update IN with IN' after the check of IN and IN'
    until maxIter or avgDifference(IN,IN') ≤ smallDiff
(9) rankedInterests=Sort words in decreasing order of IN
    return the N rankedInterests with highest scores
```

Figure 1. Determining readers' words of interest.

Two semantic features are used in PageRank. Firstly, we weigh edges according to connecting words' syntactic parts-of-speech via edge multiplier *m*. We distinguish content words (e.g., nouns, verbs, adjectives and adverbs) from are not and implement three different levels of content-word score aggregation. Particularly, we have *slightly* content word centered score propagation when $m>1$ in Step (4a) and $m=1$ in Step (4b) to (4d), while we have *moderate* content word aggregation when $m>1$ in Step (4a) and (4d) and $m=1$ in Step (4b) and (4c). The third is to *aggressively* make a non-content word's score flow to its content word partners by setting *m* in Step (4a) and $1/m$ in Step (4b) where $m>1$, and, circulate more w_i's score to content words if w_i is a content word (i.e., $m>1$ in Step (4c) and (4d)). The second semantic feature concerns source of words. Words may come from authors or readers, and *srcWeight* is set to α if *st* is from *ART* and $1-\alpha$ otherwise. Smaller α's favor readers' perspectives more while functioning as a PageRank keyword extraction system if α is one.

We set the one-by-*v* matrix **IP** of interest preference model using interest preferences for words in Step (6) and initialize the matrix **IN** of PageRank scores. Here we use word interestingness scores in Step (7). Then we re-distribute words' interestingness scores until the number of iterations or the average score differences of two consecutive iterations reach their respective limits. In each iteration, a word's interestingness score is the linear combination of its interest preference score and the sum of the propagation of its inbound words' previous PageRank scores. For the word w_j and any edge (w_i,w_j) in *ART* and any edge (w_k,w_j) in *qualityFB*, its new PageRank score is computed as

$$\mathbf{IN}'[1,j]=\lambda\times\begin{pmatrix}\alpha\times\sum_{i\in v}\mathbf{IN}[1,i]\times\mathbf{EW}[i,j]+\\(1-\alpha)\times\sum_{k\in v}\mathbf{IN}[1,k]\times\mathbf{EW}[k,j]\end{pmatrix}+(1-\lambda)\times\mathbf{IP}[1,j]$$

Once the iterative process stops, we rank words according to their final interestingness scores and return *N* top-ranked words.

4 Experiments

In this section, we first present the data sets for training and evaluating *InterestFinder* (Section 4.1). Then, Section 4.2 reports the experimental results under different window sizes, content-word aggregation levels, estimation strategies of interest preferences.

4.1 Data Sets

We collected 6,600 articles from the blog website Wretch (www.wretch.cc) in November, 2012. In total, there were twelve first-level topics and 45 categories at the second tier. The example pre-defined two- to three-tier topic ontology ranged from Travel:Domestic to Life:Pets or from Fashion:Makeup to Techonology:Games. Author-specified topic information was exploited to derive the estimation scores of interest preferences in Section 3.2. We also collected readers' feedback to the articles. We randomly chose 30 articles from training set for testing. Two human

judges annotated interested words after reading the articles in the test set.

	nDCG	P	MRR
w/o	.778	.397	.728
agr@m=2	.765	.390	.719
mod@m=2	.782	.390	.747
slg@m=2	**.792**	**.397**	.741

Table 1. System performance of different content-word aggregation levels at N=5.

4.2 Experimental Results

Our evaluation metrics are normalized discounted cumulative gain nDCG (Jarvelin and Kekalainen, 2002), precision (i.e., P), and mean reciprocal rank (i.e., MRR). We first examine the effectiveness of our semantic feature regarding content words in interest predictions. Table 1 suggests that while *slight* (*slg*) content word propagation is helpful, *moderate* (*mod*) and *aggressive* (*agr*) are not. Inflating content words' statistics is simply sufficient. In addition, we found that smaller window size (WS=3) fit more to our context of mixed-code blogs, while suitable window sizes were much larger in news articles and research abstracts (Liu et al., 2010).

Table 2 summarizes the interest prediction quality of two baselines, *entropy* and *tfidf*, and PageRank (PR) with different interest preference estimations on test set. In Table 2, *entropy* and *tfidf*, taking local (the article) and global (whole article collection) information into account, outperform PageRank using solely local information (*PR+tf*). Among all, *PR+tfidf* achieves the best performance. Compared to *PR+Pr*'s, entropy in *PR+PrEntropy*'s does help to discern topical interest words. Moreover, the benefit of entropy is more evident when better estimation strategy *Pr(tp|w)* is applied: common words receive too much attention in *Pr(w|tp)* making readers' interest words harder to come by.

(a) @N=5	nDCG	P	MRR
Entropy	.677	.287	.659
Tfidf	.719	.313	.676
PR+tf	.657	.310	.632
PR+Pr(w|tp)	.631	.290	.583
PR+Pr(tp|w)	.673	.317	.639
PR+PrEntropy(w|tp)	.636	.283	.584
PR+PrEntropy(tp|w)	**.773**	**.337**	**.725**
PR+tfidf	**.792**	**.397**	**.741**

Table 2. System performance using article information alone at N=5.

We further exploit the collected reader feedback to train the baseline *tfidf* and our best system *PR+tfidf*. Table 3 compares their interest predictions against judges' interest and annotated words, within reader feedback, of interest in the articles. Note that the *tfidf* on reader feedback alone does not perform better.

(a) @N=5	judges' interest	general readers' interest		
	nDCG	hit	nDCG	MRR
$(tfidf)_{none}$	.719	.10	.087	.075
$(tfidf)_{all}$	.699	.10	.079	.072
$(PR+tfidf)_{none}$	.792	.19	.137	.122
$(PR+tfidf)_{Coverage}$	.805	**.30**	**.186**	**.166**
$(PR+tfidf)_{Focus}$	.779	.27	.156	.137
$(PR+tfidf)_{Coverage+Focus}$	.794	**.30**	**.182**	**.164**

Table 3. System performance using *slg* at m=4, WS=3, α=0.4 and N=5

In Table 3 we observe that (1) using all reader feedback is no better than using none (rows of *tfidf*) because not all feedback respond to the articles; (2) semantic feature of content source works well with *Coverage*- and *Focus-evaluate*. And *Coverage*- and *Focus-evaluate* are effective in checking informativity of social interaction data. $(PR+tfidf)_{Coverage}$ or $(PR+tfidf)_{Focus}$ achieves better performance on general readers' interest while maintaining the prediction power on judges' interest. (3) the chain of *Coverage*- and *Focus-evaluate* $(PR+tfidf)_{Coverage+Focus}$ further prunes 6 and 12 percent of the reader sentences compared to the individual, and, encouragingly, using one-fourth of reader interactions still helps.

Based on the findings in Table 2 and 3, we believe that proposed interest preference models, semantic features (i.e, content source and content word), and the informativity check on social interaction content are simple yet helpful in suggesting good and representative reader interests.

5 Conclusion

We have introduced a method for predicting reader interest in an article. In interest prediction, we turn to social interaction content instead of reader profile and browse history. The method involves estimating topical interest preferences, screening public reader responses, and leveraging semantic features such as words' sources (i.e., from article authors or readers) and words' parts-of-speech in PageRank. We have implemented and evaluated the method as applied to interest analysis. In two separate evaluations, we have shown that *quality* social interaction content and semantic aware PageRank help to accurately cover broader spectrum of reader interest.

Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC101-2628-E-224-001-MY3.

References

Scott A. Golder and Bernardo A. Huberman. 2006. Usage patterns of collaborative tagging systems. *Information Science*, 32(2): 198-208.

Harry Halpin, Valentin Robu, and Hana Shepherd. 2007. The complex dynamics of collaborative tagging. In *Proceedings of the WWW*, pages 211-220.

Kalervo Jarvelin and Jaana Kekalainen. 2002. Cumulated gain-based evaluation of IR technologies. *ACM Transactions on Information Systems*, 20(4): 422-446.

Hongxia Jin. 2012. Content recommendation for attention management in unified social messaging. In *Proceedings of the AAAI*, pages 627-633.

Quanzhi Li, Yi-Fang Wu, Razvan Bot, and Xin Chen. 2004. Incorporating document keyphrases in search results. In *Proceedings of the Americas Conference on Information Systems*.

Zhenhui Li, Ging Zhou, Yun-Fang Juan, and Jiawei Han. 2010. Keyword extraction for social snippets. In *Proceedings of the WWW*, pages 1143-1144.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the EMNLP*, pages 366-376.

Chris D. Manning and Hinrich Schutze. 2000. *Foundations of statistical natural language processing*. MIT Press.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing orders into texts. In *Proceedings of the EMNLP*, pages 404-411.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, pages 311-318.

Manos Tsagkias and Roi Blanco. 2012. Language intent models for inferring user browsing behavior. In *Proceedings of the SIGIR*, pages 335-344.

Ryen W. White, Peter Bailey, and Liwei Chen. 2009. Predicting user interest from contextual information. In *Proceedings of the SIGIR*, pages 363-370.

Wei Wu, Bin Zhang, and Mari Ostendorf. 2010. Automatic generation of personalized annotation tags for Twitter users. In *Proceedings of the NAACL*, pages 689-692.

Mao Ye, Xingjie Liu, and Wang-Chien Lee. 2012. Exploring social influence for recommendation- a generative model approach. In *Proceedings of the SIGIR*, pages 671-680.

Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. 2011. Topical keyword extraction from Twitter. In *Proceedings of the ACL*, pages 379-388.