2013 IEEE 13th International Conference on Data Mining Workshops

Interest Analysis using Semantic PageRank and Social Interaction Content

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

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

***Abstract*—Social media has long been a popular resource for sentiment analysis and data mining. In this paper, we learn to predict reader interest after article reading using social interaction content in social media. The abundant interaction content (e.g., reader feedback) aims to replace typically private reader profile and browse history. Our method involves estimating interest preferences with respect to article topics and identifying quality social content concerning informativity. During interest analysis, we combine and transform articles and their reader responses into PageRank word graph to balance author- and reader-end influence. Semantic features of words, such as their content sources (authors vs. readers), syntactic parts-of-speech, and degrees of references (i.e., significances) among authors and readers, are used to weight PageRank word graph. We present the prototype system, *InterestFinder*, that applies the method to reader interest prediction by calculating word interestingness scores. Two sets of evaluation show that traditional, local PageRank can more accurately cover more span of reader interest with the help of topical interest preferences learned globally, word nodes' semantic information, and, most important of all, quality social interaction content such as reader feedback.**

Keywords—interest analysis; social interaction content; PageRank; social media; reader feedback; interest preferences

I. INTRODUCTION

Many natural language texts such as news and research articles, blog and micro-blog posts, and updates in social networking are generated on the Web every day, and an increasing number of Web services target extracting keywords, mining opinions, and tracking events in these texts. Some services may even recommend article contents for readers.

Traditional keyword extraction tools such as KEA (www.nzdl.org/Kea/) typically look at texts from authors' perspective and calculate the importance of a word in the texts themselves. However, keywords obtained this way are not necessarily words that interest readers or words that catch readers' eyes. These texts could be analyzed more towards readers' side if a system exploited social interaction content (e.g., reader feedback) in social media.

In this paper, we predict the topic words in texts that catch readers' attention, or readers find interesting. We do not pay attention to the sentiment polarities (i.e., negativity or positivity) that readers hold for these words. We focus on identifying interesting topic words within texts, since words may be omitted and these implicit interest words are more than challenging (implicit expressions are also considered challenging in sentiment analysis [1]).

Consider the Web post in Figure 1 as our example of interesting topic word prediction or, in a sense, interest analysis. The post describes a newly-renovated ancient house and the history, life style, and surrounding sightseeing sites of the historical city where the house is located. Most keyword tools can easily extract key words *the old house* (*謝宅*) and *the historical city* (*台南*). But based on the topics discussed in the reader feedback, readers are also interested in the post's less-frequent word *life style* (*生活*), *traditional market* (*市場*) and single-occurrence *rental fees* (*費用*). Intuitively, articles' social interaction content from readers can be accumulated to represent readers' viewpoints and their browsing habits. And this content can bias a keyword extractor towards an interest predictor even without readily available reader profile and browse history.

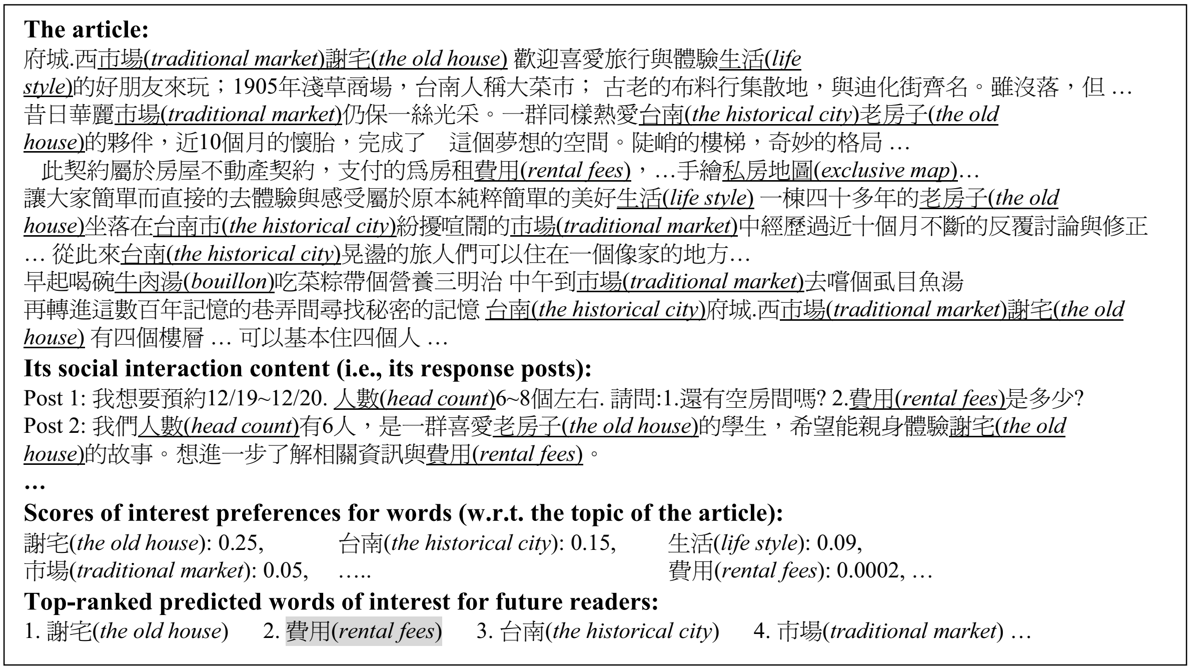
We present a new system, *InterestFinder*, that learns to profile an article in terms of reader interests. An example *InterestFinder* interest profile of a Web article is shown in Figure 1. *InterestFinder* has determined the scores of interest preferences for words in the article with respect to the article topic. *InterestFinder* learns these topic-related scores automatically during training by analyzing a collection of articles. We will describe the *InterestFinder* training process in more detail in Section III.

At run time, *InterestFinder* exploits PageRank and semantic features to find topic words of readers' interest. Specifically, it first transforms an article into a PageRank word graph. To hear readers' side of the story, *InterestFinder* combines the word graph with the one built from social interaction content. Semantic features are used to weigh and distinguish word nodes in PageRank including words' origin (i.e., article or reader feedback), parts-of-speech, and reference distribution among an article and its reader feedback. Finally, *InterestFinder* iterates with word interestingness scores to find interest terms. In Figure 1 we can see that the topic word *rental fees* (*費用*) has gained more interest (i.e., received more attention) by accommodating social interaction content.

[*] Corresponding author.

978-0-7695-5109-8/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDMW.2013.12

IEEE computer society



The article:
府城.西市場(*traditional market*)謝宅(*the old house*) 歡迎喜愛旅行與體驗生活(*life style*)的好朋友來玩；1905年淺草商場，台南人稱大菜市； 古老的布料行集散地，與迪化街齊名。雖沒落，但 …
昔日華麗市場(*traditional market*)仍保一絲光采。一群同樣熱愛台南(*the historical city*)老房子(*the old house*)的夥伴，近10個月的懷胎，完成了　這個夢想的空間。陡峭的樓梯，奇妙的格局 …
此契約屬於房屋不動產契約，支付的爲房租費用(*rental fees*)，…手繪私房地圖(*exclusive map*)…
讓大家簡單而直接的去體驗與感受屬於原本純粹簡單的美好生活(*life style*) 一棟四十多年的老房子(*the old house*)坐落在台南市(*the historical city*)紛擾喧鬧的市場(*traditional market*)中經歷過近十個月不斷的反覆討論與修正
… 從此來台南(*the historical city*)晃盪的旅人們可以住在一個像家的地方…
早起喝碗牛肉湯(*bouillon*)吃菜粽帶個營養三明治 中午到市場(*traditional market*)去嚐個虱目魚湯
再轉進這數百年記憶的巷弄間尋找秘密的記憶 台南(*the historical city*)府城.西市場(*traditional market*)謝宅(*the old house*) 有四個樓層 … 可以基本住四個人 …

Its social interaction content (i.e., its response posts):
Post 1: 我想要預約12/19~12/20. 人數(*head count*)6~8個左右. 請問:1.還有空房間嗎? 2.費用(*rental fees*)是多少?
Post 2: 我們人數(*head count*)有6人，是一群喜愛老房子(*the old house*)的學生，希望能親身體驗謝宅(*the old house*)的故事。想進一步了解相關資訊與費用(*rental fees*)。
…

Scores of interest preferences for words (w.r.t. the topic of the article):
謝宅(*the old house*): 0.25, 台南(*the historical city*): 0.15, 生活(*life style*): 0.09,
市場(*traditional market*): 0.05, ….. 費用(*rental fees*): 0.0002, …

Top-ranked predicted words of interest for future readers:
1. 謝宅(*the old house*) 2. 費用(*rental fees*) 3. 台南(*the historical city*) 4. 市場(*traditional market*) …

Figure 1. Example *InterestFinder* interest analysis for the Web article

In our prototype, *InterestFinder* returns topic words for interest evaluation; alternatively, these topic words can be used for on-topic sentiment analysis (See [1]), used as candidates for social tagging the article, or used as input to an article recommendation system.

The rest of the paper is organized as follows. We review work on keyword extraction, social tagging and content recommendation in the next section. Then we present our method for automatically estimating word interestingness scores using interest preferences for words and semantically-motivated PageRank (Section III). As part of our evaluation, we compare the interest prediction power of several baselines and our system *InterestFinder* of different settings (Section IV). Section V concludes this paper.

II. Related Work

Keyword extraction has been an area of active research. Recently, the state-of-the-art keyword extraction methods have been applied to a myriad of natural language processing tasks including document categorization and summarization [2], indexing [3], and text mining on social networking or micro-blogging services (e.g., understanding social snippets [4] and Twitter contents [5] or profiling Twitter users [6]). In our work we address an aspect of keyword extraction that focuses on reader interests. More specifically, we identify topic words that readers find interesting but ignore readers' sentiment polarities towards these words.

The body of keyword extraction systems focuses on learning word statistics in a document collection. Traditional approaches such as term frequency and inverse document frequency (i.e., tfidf), word entropy of information theory, and statistically improbable phrases (i.e., SIP), compute the distributions of words in documents (local information) and across documents (global information). On the other hand, [7] transforms word sequences into word graph and uses connectivity to extract keywords while [8] additionally considers edge type and node significance. In contrast, we extract keywords from readers' perspective. That is, we extract words that arouse readers' interest.

In studies more closely related to our work, [5] and [9] present PageRank algorithms for keyword analyses using (article) topic information. The main difference from our current work is that we integrate social content and topical interest preferences into semantic-aware PageRank. Also, we exploit author-specified article topics instead of automatic ones.

Predicting interest words, or interesting topic words, in texts can be useful for social tagging, content recommendation, and on-topic sentiment analysis [1]. We elaborate on the former two fields of research.

Collaborative tagging or social tagging describes the process where users provide metadata in the form of keywords or interest terms to the media content ([10]; [11]). The media content includes bookmarks, photographs and so on. While [10] and [11] emphasize user (tagging) activity or tag frequencies, we analyze articles and their social interaction content to predict reader interest. The returned predictions can serve as candidate tags in social tagging for understanding, learning, or navigating social content.

Recent work, on the other hand, has been done on reader profiling for content (e.g., articles or websites) recommendation. For example, [12] examines five types of contextual information (e.g., search queries) in website recommendation while [13] further explores social influence (e.g., readers' friends) on item recommendation. Moreover,

[14] concentrates on analyzing users' browsing behavior on news articles, and [15] recommends contents through a unified, personalized messaging system. Since most reader information (e.g., reader profile and browse history) is not publicly available, in this paper we accumulate social interaction content to help determine the interest of future reader. To the best of our knowledge, we are the first to evaluate the applicability of feedback content and PageRank in interest analysis.

In contrast to the previous research, we present an interest prediction system that 1) learns interest preferences with respect to domain topics, 2) determines usefulness of social interaction content and uses the content to represent readers' opinions or browsing habits, and 3) weighs PageRank word nodes considering their semantic features.

III. The IntestestFinder System

Submitting articles to keyword extraction tools for interest analysis does not work very well. Keyword tools typically look at articles from authors' perspective. Unfortunately, readers' words of interest may be ranked low by the keyword tools due to their less frequent or single appearance. To predict reader interest, a promising approach is to combine articles with their *quality* social interaction content expected to represent readers' opinions on the article.

A. Problem Statement

We focus on identifying a set of topic words within an article that are likely to interest readers or catch readers' eyes. These words are then returned as the article's interest predictions for future readers. The returned words can be examined directly, used as candidates for social-tagging the article, incorporated into on-topic sentiment analysis [1], or passed on to article recommendation systems for article retrieval. Thus, it is crucial that a reader interest be present in this set of predicted interest words. At the same time, the set of interest predictions cannot be so large that it overwhelms readers or the subsequent (typically computationally expensive) systems. Therefore, our goal is to return a reasonable-sized set of topic words that, at the same time, contain most readers' interests after reading the article. We now formally state the problem that we are addressing.

Problem Statement: We are given an article collection of various domain topics in social media (e.g., blogs), an article *ART*, and its social interaction content (e.g., reader feedback) *FB*. Our goal is to determine a set of topic words that are likely to contain an interest of future readers after reading *ART*. For this, we combine *ART* with *quality* feedback from *FB*, and view words in the sense of interestingness w.r.t. the *ART*'s topic, such that the top-ranked *N* interesting words are likely to cover most readers' interests in *ART*.

In the rest of this section, we describe our solution to this problem. First, we define strategies for estimating interest preferences under different article topics (Section III-B). These strategies rely on a set of article-topic pairs collected from the Web (Section IV-A). Finally, we show how *InterestFinder* predicts reader interest leveraging informative social responses and semantic features in PageRank (Section III-C).

B. Estimating Topical Interest Preferences

We attempt to estimate interest preferences with respect to a wide range of article topics. Basically, the estimation is to calculate the significance of a word in a domain topic. Our learning process is shown in Figure 2.

(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 various strategies
(4) Output word-and-interest-preference-score pairs for various strategies

Figure 2. Outline of the learning process.

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 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 articles' topics are specified by authors themselves and the *word* mentioned here may come from articles alone or articles together with their social interaction content (See Section IV).

The third stage involves estimating interest preferences for words across articles and across domain topics using sets of (*art*,*w*)'s and (*tp*,*w*)'s. In our paper, four popular estimation strategies in Information Retrieval and two extensions are implemented and compared. They are as follows.

- **tfidf**. $tfidf(w)=freq(art,w)/appr(art',w)$ where term frequency (w) in an article is divided by its appearance in the article collection to distinguish interesting words from common words.
- **Pr(w|tp)**. $Pr(w|tp)=freq(tp,w)/\sum_{w'} freq(tp,w')$ where a word's Maximum Likelihood Estimation of a given topic is calculated to reflect a word's interestingness or significance in the topic.
- **Pr(tp|w)**. $Pr(tp|w)=freq(tp,w)/\sum_{tp'} freq(tp',w)$ where topic-wise word senses of a word is computed to indicate topic relatedness.
- **entropy**. $\text{entropy}(w)= -\sum_{tp'} Pr(tp'|w)\times\log(Pr(tp'|w))$ where a word's uncertainty in topics is used to estimate its topic spectrum or its focus on domain topics.
- **Pr-Entropy(w|tp)**. This estimate further considers topic uncertainty in MLE, that is, $Pr(w|tp)/2^{\text{entropy}(w)}$.
- **Pr-Entropy(tp|w)**. Similarly, the last estimate incorporates entropy of information theory into topic-wise word senses, that is, $Pr(tp|w)/2^{\text{entropy}(w)}$.

Notice that these six estimations take global information (i.e., article collection) into account and will be used in PageRank which inter-connects words locally (i.e., within an article).

C. Predicting Interests for Future Reader

Once topical interest preferences for words are learned, *InterestFinder* then predicts reader interest using the procedure in Figure 3. In this procedure we exploit social interaction content and PageRank nodes' semantic features to identify the interesting topic words, or predict future readers' interests in articles. We pay no attention to readers' sentiment polarities towards these words and we do not use reader-end information such as reader profile and browse history along the process of interest prediction. Figure 4 visualizes Figure 3 in a way.

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

Figure 3. Determining readers' words of interest.

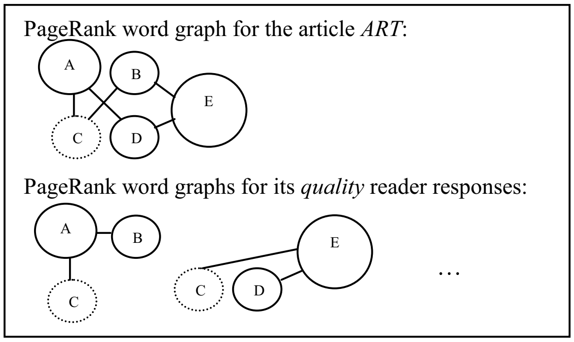

PageRank word graph for the article ART:
A
B
C
D
E
PageRank word graphs for its quality reader responses:
A
B
C
E
C
D
…

Figure 4. Word graph visualization. Note that circles denote word nodes in texts, dotted circles denote nodes of non-content words, and circles' dimensions denote their word references among readers (where the bigger the circle, more reference the word has).

```
procedure identifyInformativeFB($ART$,$FB$,$IntPrefs$)
(1) $ngrams_{art}$=generateNgram($ART$)
(2) $Focused$=findFocused($IntPrefs$)
(3) $selectedSt$=NULL
     for each response $rp$ in $FB$
       for each sentence $st$ in $rp$
(4a)     $ngrams_{st}$=generateNgram($st$)
(4b)     informativity$_{co}$=Coverage-evaluate($ngrams_{st}$,$ngrams_{art}$)
(4c)     informativity$_{fo}$=Focus-evaluate($ngrams_{st}$,$Focused$)
(4d)     append $st$ into $selectedSt$ if conditions hold
     return $selectedSt$
```

Figure 5. Identifying quality reader responses.

As one may speculate, not all social interaction content responds to the article. Take the scenario in Figure 1 for instance. Some readers show their likes/dislikes about the article, some express their thinking or ask questions on the article topics, and others respond with "less informative" messages such as commercial advertisements. Considering all social content may degrade the system performance (It does. See Section IV). As a result, we screen reader feedback in Step (1) based on the article *ART*, its feedback *FB* and interest preference scores *IntPrefs*.

Figure 5 shows the algorithm for identifying *quality* reader responses in social media. In this algorithm, each response is evaluated at sentence level concerning informativity. And we check informativity in the following two aspects.

The first concerns the topic cohesion between a reader response sentence *st* and the article *ART*. Inspired by BLEU's [16] weighted ngram precision in machine translation, we compute the weighted ngram coverage of *st* (Step (4b) in Figure 5) over *ART*. And we favor the coverage of longer ngrams.

The second, on the other hand, considers the topic distributions of words in *st*. We first rank and identify the words expected to have more focused topics in nature (i.e., low topic uncertainty). Entropy estimation in Section III-B is used for this purpose to find *Focused* (Step (2)). Afterwards, the informativity on topic focus of *st* is computed as the percentage of its words appearing 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 finding quality feedback *qualityFB*, Figure 3 further constructs a word graph for both the article and its quality social content (one can also think of this as combining word graphs from *ART* and its *qualityFB*, visualized in Figure 4). 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 (3) and (4)). Note that the graph is directional (pointing from w_i to w_j; though Figure 4 indicates otherwise) and that edge weights are the words' co-occurrence counts satisfying window size limit *WS*.

In this paper, semantic features of word nodes are used to make PageRank semantic-aware. We use three types of semantic information which are elaborated as below.

First, we weigh edges according to the connecting word nodes via edge multiplier *m*. Three different levels of weighting mechanisms concerning content words are implemented. Content words are nouns, verbs, adjectives and adverbs. For the level of *slightly* content word centered, we set m>1 in Step (3a) and m=1 in Steps (3b) to (3d). We set m>1 in Step (3a) and (3d) and m=1 in Step (3b) and (3c) for *moderately* content word centered mechanism. Or we may *aggressively* propagate more words' scores to their connecting content word by setting m>1 in Steps (3a) to (3d).

The second semantic feature takes the origin of the sentence into account. We weigh the content source from *ART* and *qualityFB* accordingly using *srcWeight* (Step (3)): *srcWeight* is set to α if *st* is from *ART* and 1-α otherwise. We use α to make sure both authors' and readers' voice are heard and to bias our interest analysis. Smaller α's favor readers' perspectives more.

We exploit word nodes' reference distributions across the article and its reader responses as the third semantic feature (Step (5)). Intuitively, if a word is referred by the author and most of the readers, it is more likely to be a reader's interest. As a result, the *RD* of a word w_{i} is its appearance in the reader responses divided by the total number of the reader responses, counting the article as "one reader response." Since the degrees of reference are used to distinguish words with different levels of significance, we add one to the *RD* ratio.

Afterwards, Step (6) sets the one-by-v matrix **IP** of interest preference model using interest preferences for words. And Step (7) initializes the matrix **IN** of PageRank scores or, in our case, word interestingness scores. 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. And the sum of the propagation is further weighted by the word's degree of references. Specifically, for the word w_{j}, its edge (w_{i},w_{j}) in *ART*, and its edge (w_{k},w_{j}) in *qualityFB*, its PageRank score is computed as

$$\begin{aligned}\mathbf{IN'}[1,j] = \lambda \times \{\, \alpha \times \textstyle\sum_{i\epsilon v} \mathbf{IN}[1,i] \times \mathbf{EW}[i,j] \times \mathbf{NS}\,[j,j] \\ + (1\text{-}\alpha) \times \textstyle\sum_{k\epsilon v} \mathbf{IN}[1,k] \times \mathbf{EW}[k,j] \times \mathbf{NS}\,[j,j] \,\} \\ + (1\text{-}\,\lambda\,) \times \mathbf{IP}[1,j].\end{aligned}$$

In the end, we rank words according to their final interestingness scores and return N top-ranked words as interesting topic words or interest predictions for future readers of the given article. An example interest analysis for a Web article on our working prototype is shown in Figure 1. Note that the article's single-appearance word *rental fees* (*費用*) has received more attention in interestingness by exploiting social interaction content in social media.

IV. Experiments

InterestFinder was designed to identify topic words of reader interest in an article using social interaction content. As such, it will be trained and evaluated over the articles in social media. Furthermore, since the goal of *InterestFinder* is to predict a good, representative set of interest words to cover most readers' interests, we also evaluate *InterestFinder* on the reader-end content. In this section, we first present the data sets for training and evaluating *InterestFinder* (Section IV-A). Then, Section IV-B reports the experimental results under different system settings (i.e., different window sizes, levels of content-word weighting mechanisms, and estimation strategies of interest preferences).

A. Data Sets

We collected approximately 6,600 articles from a famous blog website, www.wretch.cc, in late 2012. This site pre-defined article topics for bloggers and required posts come with topics. The used two- to three-tier topic ontology ranged from Travel:Domestic to Life:Pets, from Fashion:Makeup to Techonology:Games, or from Life:Food to Finance:Investment. In total, there were twelve first-level topics (i.e., art, travel, life, sports, entertainment, fashion, technology, learning, finance, society, family, and club) and they were further fine-grained to 45 categories at the second tier. The author-specified topic information was used to derive the estimation scores of interest preferences in Section III-B.

To hear reader side of the story, we also collect social interaction content (i.e., reader feedback) of the articles. Both social media data were segmented using CKIP Chinese segmentor [17]. These Web posts are mostly in Chinese but are sometimes mixed-code, that is, in Chinese and English at the same time.

Among these training articles we randomly chose 30 for testing. Table I summarizes the statistics of our data sets. On average, there were 18.3 and 17.6 reader responses per article in the training and test set respectively.

As for gold standard, two human judges were asked to annotate interested words in the test set. Take the Web post in Figure 1 for example. One judge annotated *the old house* (*謝宅*), *rental fees* (*費用*), and *exclusive map* (*私房地圖*) as terms of interest while the other further annotated *life style* (*生活*), *traditional market* (*市場*) and *bouillon* (*牛肉湯*).

With social interaction content at hand, on the other hand, we can evaluate *InterestFinder* on predicting interests of the responding readers. Therefore, same judges were also instructed to relate to the responding readers and identify *these readers*' terms of interest (in the articles) within their feedback. Compared to two judges' interests, this constitutes our experiment of majority readers' interest prediction (recall that each test article had 17.6 reader responses on average. The experiment would report the system performance on these 17.6 readers). Among 528 reader responses in the test data, judges respectively pinpointed 272 and 267 responses with clear readers' interested terms in the articles. In these responses, they marked 438 and 499 topic words that responding readers may find interesting. The statistics suggest not all feedback responds to the article and not all feedback contains reader interest (still, we could test our system on predicting more readers' interests, usually more than two). In fact, only half of the replies responded with interest information, and they usually responded to a topic word or two in the articles. Take the two response posts in Figure 1 for illustration. The *head* *count* (人數), *rental fees* (費用), and *the old house* (謝宅) were annotated and viewed as the responding readers' words of interest after reading the article.

Table I. Statistics of our (a) training and (b) testing data sets.

(I.a)	# words	Avg # words per article	vocabulary size
article	4,997K	757	164K
article + reader feedback	8,962K	1,358	217K
(I.b)			
article	27K	925	6K
article + reader feedback	40K	1,363	7K

Table II. System performance of different content-word weighting mechanisms @ *N*=5.

	nDCG	P	MRR
w/o	.778	.397	.728
agr@m=2	.765	.390	.719
agr@m=4	.754	.370	.707
mod@m=2	.782	.390	.747
mod@m=4	.765	.390	.719
slg@m=2	**.792**	**.397**	.741
slg@m=4	**.792**	**.397**	.741

Table III. Performance w.r.t. window sizes @ *N*=5.

	WS=2	*WS*=3	*WS*=6	*WS*=10
nDCG	.765	**.792**	.774	.733
P	.410	.397	.343	.350
MRR	.736	**.741**	.741	.686

B. Experimental Results

In this section, we report the evaluation results using the methodology and data sets described in the previous sections. And our evaluation metrics are nDCG [18] (standing for normalized discounted cumulative gain), P (for precision), and MRR (for mean reciprocal rank).

Different levels of content-word centralization are first examined in interest predictions. As Table II suggests, while *slight* (i.e., *slg*) centralization is helpful, *moderate* (i.e., *mod*) is not. The *aggressive* (i.e., *agr*) performs the worst. *agr* deflates non-content words' and inflates content words' significances by too much that it poorly reflects words' inter-connectivity, thus degrading the content word unaware PageRank (i.e., *w/o*). It seems that increasing the statistics propagation from non-content words to content words is simply sufficient.

Table III reports the performance of our *slightly* content word centered system with different window sizes. As one can see, smaller window sizes (but not too small) fit more to our context of blogosphere. This is contradictory to the findings in [9] where larger window sizes are more suitable in news articles and research abstracts. We attribute this small window effect to blogosphere's causal writing style and the language in use which obviously bond words in proximity.

Next, we compare estimation strategies for interest preferences with the current best-performing system's configuration. Table IV summarizes the interest prediction quality of our semantic-aware PageRank using different interest preference estimates (i.e., *PR+tf*, *PR+tfidf* and etc.) and two baselines (i.e., *entropy* and *tfidf*) on the test set.

Table IV. System performance (trained on articles alone) @ (a) *N*=5 (b) *N*=3 (c) *N*=1.

(IV.a)	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**

(IV.b)	nDCG	P	MRR
entropy	.667	.356	.644
tfidf	.651	.389	.638
PR+tf	.655	.350	.617
PR+Pr(w|tp)	.562	.328	.539
PR+Pr(tp|w)	.659	.350	.622
PR+PrEntropy(w|tp)	.562	.328	.539
PR+PrEntropy(tp|w)	**.757**	**.428**	**.717**
PR+tfidf	**.767**	**.506**	**.728**

(IV.c)	nDCG	P	MRR
entropy	.567	.567	.567
tfidf	**.600**	**.600**	**.600**
PR+tf	.500	.500	.500
PR+Pr(w|tp)	.467	.467	.467
PR+Pr(tp|w)	.500	.500	.500
PR+PrEntropy(w|tp)	.467	.467	.467
PR+PrEntropy(tp|w)	**.600**	**.600**	**.600**
PR+tfidf	**.600**	**.600**	**.600**

As shown in Table IV, global information (i.e., whole article collection) is also important: *entropy* and *tfidf* beats PageRank using solely local information (i.e., *PR+tf*). Topical interest preferences learned globally generally make RageRank a better interest predictor. Among all, *PR+tfidf* achieves the best performance across different *N*'s. Compared to *PR+Pr*'s, entropy in *PR+PrEntropy*'s does help to discern interest words. And the benefit of entropy is more evident when better estimation strategy, *Pr(tp|w)* in this case, is applied (common words receive too much attention in *Pr(w|tp)* making readers' interest words harder to come by).

In addition to the article content, we further incorporate social interaction content in training the baseline *tfidf* and our best system *PR+tfidf*. Table V compares their interest predictions against two judges' interests and annotated words, within reader feedback, of interest in the articles (i.e., the experiment of majority readers). Note that training *tfidf* on reader feedback alone does not perform better than the listed *tfidf*'s.

In Table V we observe that 1) using all reader feedback is no better than using none (*tfidf*+FB_{all} vs. *tfidf*+FB_{none}). The reason is probably that not all replies respond to the articles and some bring more harm than good; 2) *Coverage* and *Focus*

Table V. System performance using *slg* when *m*=4, *WS*=3, *α*=0.4, and (a) *N*=5 (b) *N*=3 (c) *N*=1

(V.a)	# sentences in *FB* used	judges' interest	general readers' interest		
		nDCG	hit rate	nDCG	MRR
$tfidf{+}FB_{none}$ (=$tfidf$)	0	.719	.10	.087	.075
$tfidf{+}FB_{all}$	1314 (=100%)	.699	.10	.079	.072
$PR{+}tfidf{+}FB_{none}$ (=$PR{+}tfidf$)	0	.792	.19	.137	.122
$PR{+}tfidf{+}FB_{Coverage}$	393 (=30%)	.803	**.34**	**.221**	**.182**
$PR{+}tfidf{+}FB_{Focus}$	476 (=36%)	.766	.28	.164	.139
$PR{+}tfidf{+}FB_{Coverage+Focus}$	321 (=24%)	.808	**.33**	**.210**	**.177**

(V.b)	# sentences in *FB* used	judges' interest	general readers' interest		
		nDCG	hit rate	nDCG	MRR
$tfidf{+}FB_{none}$ (=$tfidf$)	0	.651	.07	.061	.061
$tfidf{+}FB_{all}$	1314 (=100%)	.678	.10	.079	.072
$PR{+}tfidf{+}FB_{none}$ (=$PR{+}tfidf$)	0	.767	.18	.116	.110
$PR{+}tfidf{+}FB_{Coverage}$	393 (=30%)	.785	**.29**	**.186**	**.162**
$PR{+}tfidf{+}FB_{Focus}$	476 (=36%)	.773	.25	.132	.122
$PR{+}tfidf{+}FB_{Coverage+Focus}$	321 (=24%)	.784	**.30**	**.198**	**.171**

(V.c)	# sentences in *FB* used	judges' interest	general readers' interest		
		nDCG	hit rate	nDCG	MRR
$tfidf{+}FB_{none}$ (=$tfidf$)	0	.600	.07	.06	.06
$tfidf{+}FB_{all}$	1314 (=100%)	.600	.07	.053	.053
$PR{+}tfidf{+}FB_{none}$ (=$PR{+}tfidf$)	0	.600	.13	.096	.096
$PR{+}tfidf{+}FB_{Coverage}$	393 (=30%)	.600	.11	.101	.101
$PR{+}tfidf{+}FB_{Focus}$	476 (=36%)	.600	.14	.099	.099
$PR{+}tfidf{+}FB_{Coverage+Focus}$	321 (=24%)	.600	.11	.101	.101

check on informativity can select useful social interaction data and contribute to interest analysis. $PR{+}tfidf{+}FB_{Coverage}$ and $PR{+}tfidf{+}FB_{Focus}$ achieve much better performance on general readers' interest than $PR{+}tfidf$. To be specific, $PR{+}tfidf{+}FB_{Coverage}$ relatively increases hit rate by 240% and 79% compared to *tfidf* and $PR{+}tfidf$ at *N*=5. Obviously, good data is better than all data; 3) compared to the individual check, chaining *Coverage* and *Focus*, $FB_{Coverage+Focus}$, further prunes 6 and 12 percent of the reader sentences. And encouraging, the one-fourth of reader interaction data still helps (see $PR{+}tfidf{+}FB_{Coverage+Focus}$). It is worth mentioning that our third semantic feature alone relatively improves our best system by 13%. It seems that prediction power gains from knowing the reference distribution among authors and readers.

Based on the results in Table IV and V, we are modest to say that the proposed interest preference models like *tfidf* and *PrEntropy(tp|w)*, three semantic-related weighting mechanisms for word nodes, and the informativity check on social content are simple yet helpful in suggesting good and representative sets of reader interests, or topic words catching readers' eyes.

V. Future Work and Summary

In this paper we focus on using semantic-aware PageRank and social interaction content to predict interesting topic words in blog articles. Currently, we pay no attention to readers' sentiment polarities towards the words and the words outside the articles and their reader responses. In the future, we would like to devise a strategy to discover omitted interest words on reader-end social content for better interest analysis. Word omission happens frequently in blogosphere, especially in reader feedback (since most of the topic words are covered in the articles themselves). Also, we would like to examine the possibility of predicting interest words that are not covered in the articles. A good place to start is the heated discussed words which are likely to be questions or more-to-know on the article. Another interesting direction to explore is to examine the connection between reader sentiment and reader interest: will sentiment analysis on social interaction content help interest analysis, will interest analysis help on-topic sentiment detection [1], and will they benefit from each other. On the other hand, we would like to evaluate our system in the context of speech data such as audio transcripts, social tagging, and article recommendation.

In summary, we have proposed a work that falls under the umbrella of big social data analysis [19], specifically, reader interest analysis via PageRank and social interaction content. The method involves automatically estimating topical interest preferences for words, automatically screening public reader responses on informativity, and incorporating three semantic features, such as reference distribution, into PageRank. We have implemented and thoroughly evaluated the method as applied to reader interest prediction. In two separate evaluations, judges' and general readers' interest prediction, we have shown that social interaction content, topical interest preferences, and semantics of words' parts-of-speech, content sources, and degrees of references among readers, help to accurately cover broader spectrum of reader interest, even without the help of reader profile and browse history.

REFERENCES

[1] Erik Cambria, Björn Schuller, Yunqing Xia, Catherine Havasi. 2013. New avenues in opinion mining and sentiment. *IEEE Intelligent Systems*, 28(2):15-21.

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

[3] 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*.

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

[5] 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.

[6] 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.

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

[8] Divya Padmanabhan, Prasanna Desikan, Jaideep Srivastava, and Kashif Riaz. 2005. WICER: a weighted inter-cluster edge ranking for clustered graphs. In *Proceedings of the IEEE/WIC/ACM WI*, pages 522-528.

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

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

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

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

[13] 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.

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

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

[16] 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.

[17] Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In *Proceedings of the ACL Workshop on Chinese Language Processing*.

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

[19] Erik Cambria, Dheeraj Rajagopal, Daniel Olsher, and Dipankar Das. 2013. Big Social Data Analysis. In: *Big Data Computing* (R. Akerkar Ed.), chapter 13, pages 401-414.