Finding Keyphrases of Readers' Interest Utilizing Writers' Interest in Social Media

Lun-Wei Ku, Andy Lee, and Yan-Hua Chen

Institute of Information Science, Academia Sinica, Taipei, Taiwan
lwku@iis.sinica.edu.tw, {andycyrus,dorayhc}@gmail.com

Abstract. Suggesting further reading materials is an application of recommendation. Considering context, current systems usually rely on topic information and related materials to propose options for users, while users behavior is also commonly used if log information is involved. However, the users interests, which are aroused by the content of the current article they read instead of what they have had, are seldom detected from the context, and they are usually the motive that readers want to read more. This paper presents an approach to detect readers' interest from the current article they read and the users feedback of it. TED talks are utilized as the experimental materials. InterestFinder proposes interest keywords/keyphrases for each talk, where different kind of words and phrases are provided to it to find suitable candidate terms. Experiments show that the best setting proposed achieves a NDCG@50 0.6392, and the detail results are discussed. Results conclude that considering both words and phrases in a proper selection criteria benifits, and finding conceptual keyphrases as interest terms is necessary to further improve the system performance.

Keywords: interest analysis, reading recommendation, TED talk, keyword and keyphrase extraction.

1 Introduction

The Internet is a modern source for getting information. Because of the information explosion, how to satisfy users' information need efficiently is very important. Good search engines or information retrieval systems can find related articles by queries, but users might need a sequence of querying operations just to fulfill their information need brought up by the current article. To save time and effort and to automatically provide reading materials for readers, we propose a research topic named "readers' interest analysis" which aims to find possible further interest of readers after reading the current article. When browsing an article within a site, it is common for the site to suggest related articles that the reader might be interested in. For example, a reader of a movie review article is interested in where and when he could see this movie, and unless the review is from a professional movie website, it is often that the reader needs to search for the theater site and look up the time table manually as it may not be provided by the review article. Knowing the readers' interest, i.e., the time table

F.F.-H. Nah (Ed.): HCIB/HCII 2014, LNCS 8527, pp. 183–193, 2014.

and the theater of the movie (and there may be more interests) in the previous example, help to create a satisfying reading process dynamically.

Researchers have developed some content recommendation techniques. For example, when browsing a video on YouTube, suggestions for related videos are on the side [1]. Sometimes suggestions are generated by extracting topical keywords from the current article [2]. However, this research is different from the previous ones from four aspects:

1. Detecting interest terms instead of topical keywords.
2. Finding users' interest from the content of articles instead of the user behavior.
3. Detected interest does not need to be internal (in the current article).
4. The proposed method is not domain specific.

Previously we developed InterestFinder, a prototype system, based on PageRank, tfidf score of words, and social interaction content to extract keywords that reflect the reader's interest of the current article [3]. In this paper, we improve InterestFinder by also extracting keyphrases instead of only keywords to capture more complete concepts. For example, in the article introducing apple computers the keyphrase "apple computer" is extracted instead of the keyword "apple" or "computer". The quality of the extracting terms is also improved by considering coverage and coherence [4]. We further try to search interests from social media by the keyphrases proposed by InterestFinder from the current article and its user feedback (internal keyphrases). The resulting articles will be sent to InterestFinder to find more interest keyphrases (external keyphrases). These keyphrases can be utilized as the input to an article recommendation system. The system flow is shown in Figure 1.

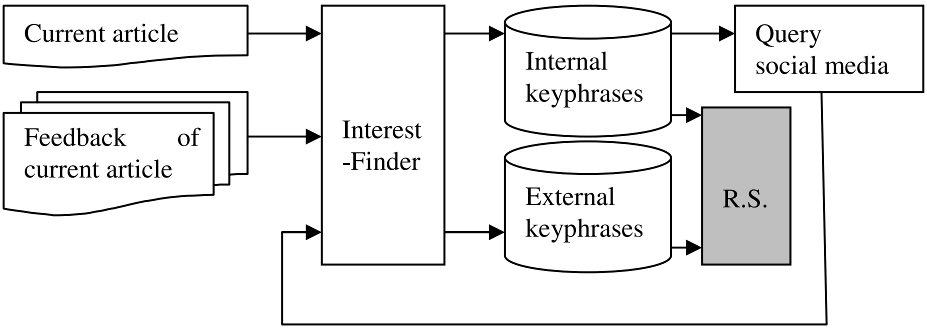

Fig. 1. System flow. R.S. is a recommendation system, which is not included in this research

2 Related Work

Keyword extraction has been a popular research area. Natural language processing tasks including document categorization and summarization [5], indexing [6], information retrieval, and text mining on social networking or micro-blogging services [7][8][9] have utilized keyword extraction techniques as a tool widely and explored them intensively. Like these researches but more specifically, this paper focuses on extracting those keywords that are reader interests, and further, not only keywords but also keyphrases. Extracting interest keyphrases is the first step toward article recommendation, i.e., reader interest prediction after article reading.

The core of keyword extraction systems mainly depends on automatically learning word statistics in a document collection. Traditional approaches such as term frequency and inverse document frequency (tfidf) poses a strong baseline. Other interesting approaches such as the word graph presented by Mihalcea and Tarau [10], which uses connectivity of words in local document to extract keywords, and word graph by Liu's research group [11], which extracts keyphrases by graph-based ranking methods, all remind us that relations between words are useful clues for keyword extraction. Therefore in our core algorithm (InterestFinder), a semantic aware PageRank is designed to gives partial scores to the candidate terms when selecting interest keyphrases.

Recently, collaborative tagging or social tagging has grown in popularity among Web services and received much attention [12][13]. In their research, user (tagging) activity or tag frequencies were analyzed. In our research, we viewed the social interaction content as a kind of social tagging which points out terms in the main article that are more likely to be interests.

HEADY proposed by Alfonseca and his team [14], which generated the headline of news articles, is another concept to find keywords/keyphrases. It is an abstraction process for the original article. It considered the people, events, time, location, objects. Though HEADY focused on the generation process, the concept that parts of speeches are utilized to select useful terms is also adopted in this paper.

3 Approach, Materials and Settings

We utilized the materials from TED talks (http://www.ted.com) for experiments. The title, English transcript and user responses of each talk were extracted. These materials were then parsed by the Stanford parser and stemmed by the Porter stemmer. Experiments were performed with two sets of talks: one set includes all talks from TED (ALL), and the other set includes 500 talks of TED topic "technology" (TECH).

By experimenting by the above different sets of terms, we expected to filter out noises and find more proper terms for extracting representative interest keywords/keyphrases. As mentioned, these candidate interest keyword/keyphrases and their ranks were given by InterestFinder [3].

Interest Finder is a system which proposes terms to indicate readers' interest by exploiting social interaction content (e.g., reader responses) and words' semantic features (e.g., content sources and parts of speech). The approach adopted by Interest Finder involves estimating topical interest preferences and determining the informativity between articles and their social content. Interst Finder considers quality responses which represent readers' opinions to balance authors' statements. The topical interest preferences were estimated by TFIDF, a traditional yet powerful measure shown in formula (1). Then semantic aware PageRank in formula (2) is used on candidates to find reader interest with the help of their interestingness scores. 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) in Fig. 2). Note that the graph is directional (pointing from w_i to w_j) and that edge weights are the words' co-occurrence counts satisfying window size limit *WS*. 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 or, in our case, word interestingness scores in Step (7). Previous experiments has showed that the social interaction content and its proposed selection process help to accurately cover more span of reader interest [3]. InterestFinder works on words. However, compared to words, phrases usually represent more complete concepts. Therefore, in our experiments, we aim to find a

```
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 IsCont-
Word(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
```

Fig. 2. Determining readers' words of interest by semantic aware PageRank

```
procedure selectInformativeFB(ART,FB,IntPrefs)
(1) ngrams_art=generateNgram(ART)
(2) Focused=findFocused(IntPrefs)
(3) selectedSt=NULL
for each sentence st in FB
(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
```

Fig. 3. Identifying quality reader responses

proper way to include phrases into the candidates for InterestFinder in order to propose better keywords/keyphrases which indicate readers' interest.

$$\text{tfidf}\left(art, w\right) = \text{freq}\left(art, w\right) / \text{artFreq}\left(w\right) \tag{1}$$

$$\mathbf{IN}'[1,j]=\lambda\times\begin{pmatrix}\alpha\times\sum_{i\in v}\mathbf{IN}[1,i]\times\mathbf{EW}[i,j]+\\(1-\alpha)\times\sum_{k\in v}\mathbf{IN}[1,k]\times\mathbf{EW}[k,j]\end{pmatrix}+(1-\lambda)\times\mathbf{IP}[1,j] \tag{2}$$

We tried not to limit the source of interest words to the current talk and its responses. (i.e., internal keyphrases in Fig.1) To find other keywords/keyphrases (i.e., external keyphrases shown in Fig.1) to indicate possible readers' interest not mentioned in the talk and responses, we searched from blogs and social media. Posts from blogs and social media serve as an bridge to link other possible related interests to the current article. It is common when people are interested in something new but not familiar with it, they may consult their friends who has been long time paying attention to it and try to know what they should start with. The idea is like that. We find people of the same interest group in social media from their posts which contain internal interest terms. Then we try to extract other interests mentioned by them. This is a process to find what "people who are interested in the internal interest terms are also interested in." In this paper Engaget serves as the external source. These extracted interests are then treated as external interest keyphrases and will be proposed as the interest keyphrases together with the internal ones.

In transcripts of talks, seven different sets of terms were considered as candidates of interest keywords and the performance of using these seven sets were compared in our experiments. These seven sets were composite of terms which are:

1. noun phrases and verb phrases: $(\text{NP+VP})_{\text{tech}}$
2. noun phrases and verb phrases (NP+VP) which contain words of parts of speech NN, NN, NNP, NNS, NNPS: $(\text{PN})_{\text{tech}}$

3. noun phrases and verb phrases (NP+VP) which contain words of parts of speech NN, NN, NNP, NNS, NNPS and other words (which are not phrases) of parts of speech NN, NN, NNP, NNS, NNPS: $(PN+WN)_{tech}$
4. words: $(W)_{tech}$
5. interest terms from (3) which contains words from (4), ranked by (3) first then (4): $(PN+WN+W)_{tech}$
6. using top 10 terms from (3) as the queries to Engaget, and find relevant posts, including post titles, post bodies and readers' feedback. For each post, calculate tf-idf scores of terms in these posts. Merge all posts and calculate the pagerank scores of composite terms. Select interest terms according to their pagerank scores and then tf-idf scores. Terms from Engaget are external keyphrases in Fig. 1: $(EXT)_{tech}$

Scores of terms for the above (a) to (f) are calculated from the talk set TECH, and those for the following

7. noun phrases and verb phrases (NP+VP) which contain words of parts of speech NN, NN, NNP, NNS, NNPS and other words (which are not phrases) of parts of speech NN, NN, NNP, NNS, NNPS: $(PN+WN)_{all}$

are from the talk set ALL.

To evaluate, three annotators annotates the interest keywords/keyphrases from the list proposed by InterestFinder. We treat these annotations as gold standard and calculate the number of interest terms and NDCG (normalized discounted cumulative gain) to know the characteristics of this research problem and the ability of the proposed approach to rank the interest keywords/keyphrases.

4 Experiment Results and Discussions

We randomly select 10 TED talks for evaluation, shown in Table 1. After extracting all interet keyphrases from these talks, we ask annotators to label their interest from top 50 keyphrases. Sample top 10 keyphrases of one of these talks, "Shyam Sankar: The rise of human-computer cooperation", are shown in Table 2. (Talk 1, http://www.ted.com/talks/shyam_sankar_the_rise_of_human_computer_cooperation.html)

Table 1. Ten talks for evaluation

Talk	Title
1	The Rise of Human-Computer Cooperation
2	The Case for Anonymity Online
3	One Laptop per Child, Two Years on
4	Reach into the Computer and Grab a Pixel
5	The Astounding Athletic Power of Quadcopters
6	One Very Dry Demo
7	My Radical Plan for Small Nuclear Fission
8	10 Top Time-saving Tech Tips
9	Why Google Glass?
10	Hack a Banana, Make a Keyboard!

Table 2. Sample interest keyphrases of Talk 1

Rank	Keyphrase	Rank	Keyphrase
1	man and machine	6	Tim Huang
2	machine	7	human-computer symbiosis
3	foreign fighter	8	A computer science titan
4	computer	9	Licklid
5	Human	10	Cooper

The numbers of the interest keyphrases labeled by the annotators from the top 50 candidates proposed by the setting (PN+WN) all are listed in Table 3. We found that the number of interest keyphrases varies among annotators and also among talks. The former conforms to our expectation as people have different interessts, and the latter tells us words and phrases in some topics are especially not proper to represent related interests.

Table 3. Number of labeled interest keyphrases@50

Annotator/Talk	1	2	3	4	5	6	7	8	9	10	AVG
A	5	6	5	10	10	5	10	7	6	5	6.9
B	11	3	4	5	5	11	9	3	4	2	5.7
C	12	12	13	14	13	4	8	3	6	3	8.8
AVG	9.3	7.0	7.3	9.7	9.3	6.7	9.0	4.3	5.3	3.3	

Table 4, 5, and 6 show the NDCG@10, 20, 50 of 10 talks from 3 annotators, respectively. The performances of 7 settings in three tables show the same tendency. From experiment results of (W)tech, we find that some words do serve as good interest terms. However, we also find that sometimes words cannot express a complete interest concept. (NP+VP)tech and (PN)tech attempt to find interest terms from phrases. (PN)tech uses a subset of candidates from (NP+VP)tech by applying a more strict selection criteria to noun and verb phrases. Experiments show that (PN)tech performs better than (NP+VP)tech as (PN)tech filters out phrases including pronouns, which are less possible to be interests.

As words and phrases may serve as good candidates of interest keywords/keyphrases, selecting interest terms from both of them is our next move. $(PN+WN)_{tech}$ takes both words and phrases which satisfy some part of speech requirements into consideration. However, compared to $(PN)_{tech}$, the performance of $(PN+WN)_{tech}$ decreases a bit, which tells that WN is not a good candidate word set for selection, at least, not good enough. Therefore, we additionally add W into $(PN+WN)_{tech}$, i.e., $(PN+WN+W)_{tech}$, to keep the candidate phrase set but enlarge the candidate word set. Experiments show that $(PN+WN+W)_{tech}$ outperforms $(PN+WN)_{tech}$ but only comparable to $(W)_{tech}$. From results of annotator A, B, and C, we further find that though performances of $(PN+WN+W)_{tech}$ and $(W)_{tech}$ are comparable, obviously annotator A prefers terms proposed from $(W)_{tech}$ while annotator C prefers that from $(PN+WN+W)_{tech}$, both of which tell us using pure phrases will not bring us the best results.

$(EXT)_{tech}$ includes candidate terms from the external source: Engadget. We have expected that involving the collaborative filtering would find us better interest terms that are not include in the original talk or its users' feedback. However, topics of retrieved articles and their comments are too diverse by querying with top 10 internal interest terms. The annotators have problems to link interest terms extracted from these articles to the original talk so that they can only select some general interest terms. Moreover, the proposed terms from Engadget include some commonly seen irrelevant terms in technology related social media or blog posts, such as display, iphone, and screen. Experiments show that the performance of $(EXT)_{tech}$ drops a lot. We will need to solve the mentioned issues to well utilize external resources.

The tfidf scores are important for InterestFinder. They calculate the topical interest preference and involve in the semantic aware PageRank as described in section 3. The postulation that enlarging the corpus for the tfidf calculation may help to detect topics more accurately is confirmed by the performance of $(PN+WN)_{all}$. Using the ALL set, $(PN+WN)_{all}$ performs betther than other settings using the TECH set, which shows discriminative terms are also more likely to be interest terms.

Figure 4 shows the NDCG@50 of each talk at the setting $(PN+WN)_{all}$. The performances vary and the individual difference among talks exists. For some talks, even performances among annotators vary a lot. This is because finding interests is a subjective task. We may only expect the system to find interests which fit most people's need. However, there are some talks whose performances are comparably low for labels from all annotators, like talk 8 and talk 10. We further analyze the proposed terms of these two talks and find some interesting phenomena. For talk 8, whose topic is about "tips", it is difficult to find words or phrases to represent related interests. Words and phrases are too short to properly express tips or procedures and even they do, tips or procedures are usually not interests. In addition, talks like this could arouse more conceptual interests, such as "useful" and "helpful" that we may not find in talks. Terms for Talk 10 are of similar problem. Talk 10 describes an interesting idea and an interesting stuff, while the name of this stuff is not well written in the article. Instead, its components, functions, and usages are described in paragraphs. Therefore, the concrete interest keywords/keyphrases should be the name of this stuff but we cannot find a proper name of it from the talk; the conceptual interest terms could be "interesting", "creative" sort of words which cannot be found either like in talk 8. Table 3 also shows that the proposed interest terms of talk 8 and talk 10 are less annotated as readers' interest in average. In fact, possible conceptual interest terms like "inspiring", "amazing" are also commonly searched by TED users and they are tagged to some talks. Therefore, we will have to provide an approach which proposes the adjectives describing the overall comment, maybe sentiment, of the talk to enlarge the candidate interest term set.

Table 4. NDCG@10 of 10 talks from 3 annotators

Setting	A	B	C	AVG
$(NP+VP)_{tech}$	0.1324	0.1485	0.0927	0.1246
$(PN)_{tech}$	0.3350	0.2930	0.2574	0.2951
$(PN+WN)_{tech}$	0.3073	0.3336	0.2236	0.2882
$(W)_{tech}$	0.4904	0.4591	0.2018	0.3838
$(PN+WN+W)_{tech}$	0.3827	0.4269	0.3034	0.3710
$(EXT)_{tech}$	0.1921	0.3746	0.2348	0.2672
$(PN+WN)_{all}$	0.5260	0.3927	0.3411	**0.4199**

Table 5. NDCG@20 of 10 talks from 3 annotators

Setting	A	B	C	AVG
$(NP+VP)_{tech}$	0.2532	0.2884	0.2103	0.2506
$(PN)_{tech}$	0.4442	0.3992	0.3143	0.3859
$(PN+WN)_{tech}$	0.3915	0.3810	0.3025	0.3583
$(W)_{tech}$	0.5715	0.4853	0.2582	0.4383
$(PN+WN+W)_{tech}$	0.5084	0.4954	0.3936	0.4658
$(EXT)_{tech}$	0.2526	0.4199	0.3229	0.3318
$(PN+WN)_{all}$	0.6239	0.4629	0.4037	**0.4968**

Table 6. NDCG@50 of 10 talks from 3 annotators

Setting	A	B	C	AVG
$(NP+VP)_{tech}$	0.4558	0.4496	0.4365	0.4473
$(PN)_{tech}$	0.5860	0.5581	0.5562	0.5667
$(PN+WN)_{tech}$	0.5652	0.5374	0.5329	0.5452
$(W)_{tech}$	0.6401	0.5940	0.4966	0.5769
$(PN+WN+W)_{tech}$	0.6257	0.6017	0.5591	0.5955
$(EXT)_{tech}$	0.3828	0.5229	0.5294	0.4784
$(PN+WN)_{all}$	0.7298	0.5856	0.6021	**0.6392**

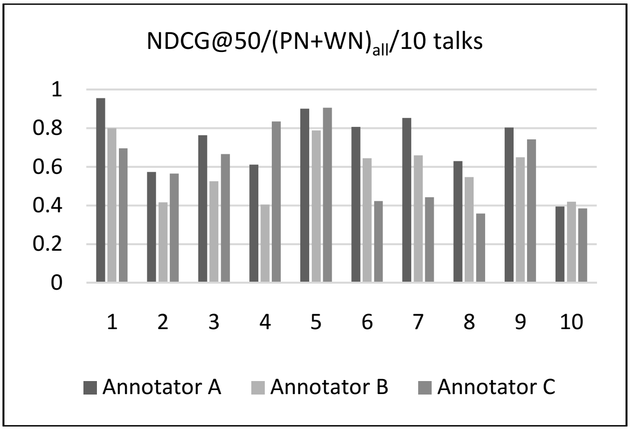

Fig. 4. NDCG@50 of $(PN+WN)_{all}$ for 10 talks

5 Conclusion and Future Work

Through the work we aimed to detect readers' interest aroused from the current article they were reading. Keywords/keyphrases were used to represent the interests for further reading materials recommendation. TED talks which usually contained novel topics were selected for experiments. We prepared several different sets of candidates for InterestFinder to extract interest terms. Experiments conclude that words and phrases could both be good terms to describe interests. However, specific words are necessary for representing a concept, while precision or relativeness is the requirement when selecting interest phrases. The proposed approach achieves an NDCG@50 of 0.6392.

We found that this research problem can actually be divided into three problems. The first is “where can we find candidate interest terms?” External sources should be considered in a suitable way. Then the second is “what kind of words/phrases can represent an interest?” After we found terms which may represent interests, we further need to solve the problem “which interest terms among them have some kind of relations with the current article?” Focusing on these three problems and solving them will be our next step to propose better interest keywords and keyphrases.

From results we know that characteristics of the articles may determine their related interest terms. Considering article types when selecting interest terms could be useful. In addition, conceptual terms which tell the overall feelings after reading articles such as “inspiring” and “funny” could also be interest terms. Integrating sentiment analysis technique and searching from the tagxonomy of talks to provide this kind of interest terms is our next goal.

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

References

1. Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., Sampath, D.: The YouTube Video Recommendation System. In: Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys 2010), pp. 293–296 (2010)
2. Phelan, O., McCarthy, K., Smyth, S.: Using Twitter to Recommend Real-Time Topical News. In: Proceedings of the third ACM Conference on Recommender systems (RecSys 2009), pp. 385–388 (2009)
3. Huang, C., Ku, L.-W.: Interest Analysis using Semantic PageRank and Social Interaction Content. In: Proceedings of the IEEE International Conference on Data Mining, SENTIRE Workshop (2013)
4. Ding, Z., Zhang, Q., Huang, X.: Keyphrase Extraction from Online News Using Binary Integer Programming. In: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), pp. 165–173 (2011)
5. Manning, C.D., Schutze, H.: Foundations of statistical natural language processing. MIT Press (2000)
6. Li, Q., Wu, Y.-F., Bot, R., Chen, X.: Incorporating Document Keyphrases in Search Results. In: Proceedings of the Americas Conference on Information Systems (2004)
7. Li, Z., Zhou, G., Juan, Y.-F., Han, J.: Keyword Extraction for Social Snippets. In: Proceedings of the WWW (WWW 2010), pp. 1143–1144 (2010)
8. Zhao, W.X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.-P., Li, X.: Topical Keyword Extraction from Twitter. In: Proceedings of the ACL (ACL 2011), pp. 379–388 (2011)

9. Wu, W., Zhang, B., Ostendorf, M.: Automatic Generation of Personalized Annotation Tags for Twitter Users. In: Proceedings of the NAACL, pp. 689–692 (2010)
10. Mihalcea, R., Tarau, P.: TextRank: Bringing Orders into Texts. In: Proceedings of the EMNLP, pp. 404–411 (2004)
11. Liu, Z., Huang, W., Zheng, Y., Sun, M.: Automatic Keyphrase Extraction via Topic Decomposition. In: Proceedings of the EMNLP, pp. 366–376 (2010)
12. Golder, S.A., Huberman, B.A.: Usage Patterns of Collaborative Tagging Systems. Information Science 32(2), 198–208 (2006)
13. Halpin, H., Robu, V., Shepherd, H.: The Complex Dynamics of Collaborative Tagging. In: Proceedings of the WWW, pp. 211–220 (2007)
14. Alfonseca, E., Pighin, D., Garrido, G.: HEADY: News Headline Abstraction Through Event Pattern Clustering. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), pp. 1243–1253 (2013)