Hindawi
Complexity
Volume 2022, Article ID 7015764, 8 pages
https://doi.org/10.1155/2022/7015764

WILEY | Hindawi

Research Article

Keyword Extraction for Medium-Sized Documents Using Corpus-Based Contextual Semantic Smoothing

Osama A. Khan,[1] **Shaukat Wasi**,[1] **Muhammad Shoaib Siddiqui**,[2] **and Asim Karim**[3]

[1]*Department of Computer Science, Mohammad Ali Jinnah University (MAJU), Karachi 75400, Pakistan*
[2]*Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia*
[3]*Department of Computer Science, Lahore University of Management Sciences (LUMS), Lahore 54792, Pakistan*

Correspondence should be addressed to Osama A. Khan; osamaahmedkhan@hotmail.com

Received 7 June 2022; Revised 2 August 2022; Accepted 9 August 2022; Published 29 September 2022

Academic Editor: Muhammad Ahmad

Keyword extraction refers to the process of selecting most significant, relevant, and descriptive terms as keywords, which are present inside a single document. Keyword extraction has major applications in the information retrieval domain, such as analysis, summarization, indexing, and search, of documents. In this paper, we present a novel supervised technique for extraction of keywords from medium-sized documents, namely Corpus-based Contextual Semantic Smoothing (CCSS). CCSS extends the concept of Contextual Semantic Smoothing (CSS), which considers term usage patterns in similar texts to improve term relevance information. We introduce four more features beyond CSS as our novel contributions in this work. We systematically compare the performance of CCSS with other techniques, when implemented over INSPEC dataset, where CCSS outperforms all state-of-the-art keyphrase extraction techniques presented in the literature.

1. Introduction

Keyword extraction can be defined as the process of selecting most significant, relevant, and descriptive terms as keywords, which are present inside a single document, where "terms" refer to distinct n-grams of any size. Keywords represent distinguished and specialized concepts and are bound to convey the informational content load of a document. Keyword extraction has major applications in the information retrieval domain, such as summarization [1, 2], indexing [3], search [4], tagging [5, 6], contextual advertising [7, 8], and personalized recommendation [9].

Documents can generally be classified into long-, medium-, and short-sized documents, where webpages, news articles, and research papers represent long-sized documents, research papers' abstracts, emails, and question-and-answer conversations characterize medium-sized documents, while microposts and Short Message Service (SMS) denote short-sized documents. Each type of document possesses unique characteristics and challenges that need to be dealt with before any keyword extraction technique can be successfully applied on it. Long-sized documents comprise large vocabulary, medium-sized documents include lack of context, while short-sized documents contain challenges related to low signal-to-noise ratio, extensive preprocessing, and multivaried text composition [10].

Replacing author-assigned keywords in research papers' abstracts, topic identification of emails, and topic recommendation for question-and-answer conversations are a few significant applications of keyword extraction from medium-sized documents in the real world.

A research paper abstract can provide a user the summary of the respective research article, in absence of his/her access to the latter. Hence, keywords extracted from research abstracts would represent the ones extracted from respective research articles. Also, research papers contain keywords that are manually tagged by respective authors. Manually tagged keywords contain bias that helps respective research papers to appear in top results, when searched by users utilizing those index terms. This can be observed by looking

at examples of ACM (https://www.acm.org/) and IEEE (https://www.ieee.org/), both of which are leading research organizations in the domains of Computer Science and Engineering, respectively, and hence possess a majority of authors in these domains, when considered together. For ACM, authors need to provide Computing Classification System (CCS) (https://dl.acm.org/ccs) Concepts that are defined by ACM, but also keywords that are defined by authors. For IEEE, authors need to provide their own defined keywords as index terms. Upon automatic selection, keywords or index terms should exclude associated bias in the search process upto a certain level, e.g., in terms of $F-$measure (see Section 4.2).

Keyword extraction has been performed in the literature on all types of documents (long-sized [11], medium-sized [12], and short-sized [13]), while utilizing various techniques. Keyword extraction techniques developed so far have been either supervised [14] or unsupervised [15]. Unsupervised techniques can be used on multiple document collections without the need for costly and time-consuming prior labeling. On the other hand, supervised techniques although require periodic training from human-labeled document collections, they still can be more accurate [16, 17].

In this paper, we present a novel supervised technique for extraction of keywords from medium-sized documents, namely Corpus-based Contextual Semantic Smoothing (CCSS). CCSS extends the concept of Contextual Semantic Smoothing (CSS) [10], which considers term usage patterns in similar texts to improve term relevance information for short-sized documents. In fact, CSS performs smoothing of the TFIDF matrix using a semantic feature, namely Phi coefficient, while keeping the corpus context into consideration. We introduce four more features beyond CSS as our novel contributions in this work in order to handle further challenges associated with medium-sized documents.

2. Related Work

PageRank is a graph-based unsupervised language-independent ranking algorithm, presented by Page et al. [18], which uses link information to iteratively assign global importance scores to webpages. PageRank is based upon the principle: "A vertex is important if there are other important vertices pointing to it," which can be regarded as voting or recommendation among vertices. In PageRank for keyword extraction, the ranking score of a candidate keyword is computed by summing up the ranking scores of all unigrams within the keyword [19–21]. Then, candidate keywords are ranked in descending order of ranking scores, and the top N candidates are selected as keywords.

Various methods have been proposed in the literature to infer latent topics of words and documents. These methods are known as latent topic models that derive latent topics from a large-scale document collection according to word occurrence information. Latent Dirichlet allocation (LDA), developed by Blei et al. [22] is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. It is representative of the latent topic models, embeds supervised learning, has more feasibility for inference, and can reduce the risk of overfitting.

Topical PageRank (TPR), proposed by Liu et al. [23], is based upon PageRank [18], which measures the importance of a word with respect to different topics. Given the topic distribution of a document, ranking scores of words are calculated with respect to those topics, and top ranked words for each topic are extracted as its keywords, thus resulting in a good coverage of the document's major topics. TPR combines the advantages of both LDA and TFIDF/PageRank, by utilizing both external topic information (like LDA) and internal document structure (like TFIDF/PageRank).

Liu et al. [24] devised an unsupervised technique for keyword extraction, which first finds exemplar terms in a document by leveraging, clustering, and semantic relatedness, which guarantees the document to be semantically covered by these exemplar terms (centroids of clusters). Then, keywords are extracted from the document using these exemplar terms. The technique incorporates term cooccurrence information and considers Noun Phrases only for keyword candidates.

Tsatsaronis et al. [25] designed SemanticRank, which is again based upon PageRank [18], but ranks both keywords and sentences in a document based on their respective relevance to the document. The technique constructs a semantic graph using terms as nodes, and their implicit links while utilizing Omiotis similarity measure, WordNet, and Wikipedia as knowledge-bases and statistical information.

3. Corpus-Based Contextual Semantic Smoothing for Medium-Sized Documents

Given a collection of medium-sized documents $D = \{d_1, d_2, \ldots, d_N\}$ and domain-specific information (stopword and standardization lists), a keyword extraction technique outputs the top K keywords from a document d_i. We divide our methodology into two phases, namely (i) keyword extraction (unigrams) and (ii) keyphrase extraction (n-grams, where $n > 1$). First, all experiments were conducted to optimize the process of keyword extraction, and then the parameters were revisited to optimize the process of keyphrase extraction.

Corpus-based contextual semantic smoothing (CCSS, see Figure 1) extends the concept of Contextual Semantic Smoothing (CSS) [10], which considers term usage patterns in similar texts to improve term relevance information. In fact, CSS performs smoothing of the TFIDF matrix using a semantic feature, namely Phi coefficient, while keeping the corpus context into consideration.

3.1. Parts of Speech Tagging. In the literature, different combinations of Parts of Speech (POS) have been employed in order to filter unlikely keywords from a document, as presented in Table 1.

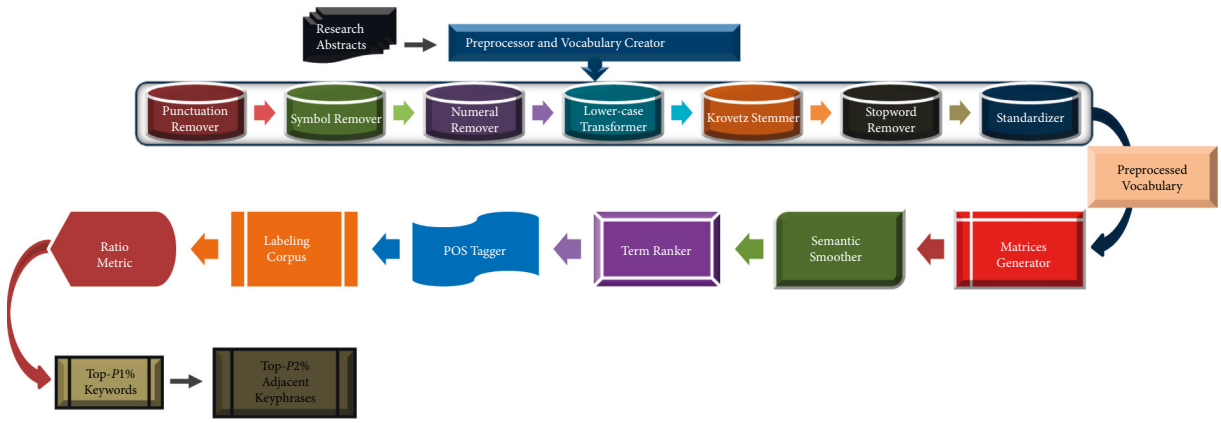

Figure 1: Corpus-based contextual semantic smoothing.

Table 1: Different combinations of POS in the literature.

S#	Reference	POS
1	[26–30]	Nouns only
2	[9, 31]	Nouns and verbs only
3	[6, 20–23, 32, 33]	Nouns and adjectives only
4	[15, 34–37]	Nouns, verbs, and adjectives only
5	[16]	Nouns, verbs, adjectives, and adverbs only
6	[38]	Verbs and adverbs only

As the first feature, we experimented with various combinations of POS (including some mentioned in Table 1), and selected the combination that considered all POS except Modal Verbs, as candidate keywords in a document. Modal Verbs are auxiliary verbs, such as "can" or "will," which are used to express modality. This combination of POS has not been used in the literature before, as evident from Table 1. Experiments related to this feature are presented in Section 5.1.

3.2. Labeling Corpus. As the second feature, we utilized a corpus consisting only of labels. We state our hypothesis as "A term should be considered as candidate keyword in a document, if it has been assigned as a label atleast once in the labeling corpus." We acquired INSPEC (https://www.theiet.org/publishing/inspec/) and ACM (https://doc.novay.nl/dsweb/Get/Document-115737/ACM-URLs.txt) collections to combine all of their labels into a single corpus. Both INSPEC and ACM collections consist of abstracts in English from scientific journal papers. Further details about the corpora are provided in Section 4.1. We experimented with various frequencies (f) of terms assigned as labels in the labeling corpus, and finally found our hypothesis to be true. In the literature, corpora have been utilized as a feature [1, 28, 39–44], however both labeling corpus in general, and this combination of corpora in particular, have not been used earlier. Experiments related to this feature are presented in Section 5.2.

3.3. Ratio Metric. As the third feature, we introduced a novel metric for each term's (t_j) eligibility for being a candidate keyword.

$$\frac{f_d}{f_l} \leq x, \tag{1}$$

where f_d represents frequency of t_j in the source document (d) under consideration, f_l represents frequency of t_j in the labeling corpus (l) under consideration, and x represents a threshold value under which the ratio of f_d and f_l should remain in order for t_j to be considered as a candidate keyword. The motivation behind developing this metric was to filter those terms as candidate keywords whose $f_d \gg f_l$. Experiments related to this feature are presented in Section 5.3.

3.4. Keyphrase Extraction. Once we had identified the significant keywords in the first phase, we moved on towards forming significant keyphrases in the second phase, through four different combinations of the two phases.

First, we considered the simplest way where all adjacently located keywords in d were utilized to form keyphrases.

Second, for all adjacently located keywords in each d, we selected the Top-P% (P is an integer between 0 and 100) keyphrases from them as significant keyphrases in order to take into consideration the varying sizes of documents.

Third, similar to selecting Top-P% keyphrases in each d as significant keyphrases, we revisited and improved the keyword extraction process by selecting Top-$P1$% ($P1$ is an integer between 0 and 100) keywords in each d as its significant keywords, and then selecting all adjacently located keywords in each d as significant keyphrases.

Fourth, we first selected Top-$P1$% (same value as resulted from the third combination of the two phases) keywords in each d as its significant keywords, and then selected Top-$P2$% ($P2$ is an integer between 0 and 100) keyphrases in each d as significant keyphrases. In the literature, keywords have been selected using the Top-P% metric; however, the process of Top-$P2$% keyphrases' selection after Top-$P1$% keywords have been selected has not been proposed earlier. Experiments related to the combinations of the two phases are presented in Section 5.4.

Table 2: Different combinations of POS.

S#		POS	Precision	Recall	F – measure
1		NO and AD only	0.3267	0.8641	0.4741
2		**NO, AD, and F only**	**0.3270**	**0.8836**	**0.4774**
3		NO, AD, and I only	0.3156	0.9176	0.4697
4	Top-N, N = *All*	NO, AD, I, and F only	0.3161	0.9370	0.4727
5		NO, AD, and V only	0.2891	0.9048	0.4382
6		NO, AD, I, V, NU, G, and AG only	0.2781	0.9631	0.4315
7		All POS except MV	0.3303	0.8195	0.4708

Bold values represent the best results achieved in terms of F-measure.

4. Data Analysis and Experimental Setup

4.1. Data Analysis. INSPEC dataset contains abstracts in English of journal papers from the disciplines of Computers and Control, and Information Technology, from 1998 to 2002, and is a collection of 2,000 documents. The keywords assigned by a professional indexer may or may not be present in the abstracts. However, the indexers had access to the full-length documents when assigning the keywords. The abstracts in this dataset contain two sections; Title and Abstract, while in this work our focus is on the Abstract section only. All experiments presented in Section 5 have been conducted over this dataset.

ACM dataset contains abstracts in English of journal, conference, and workshop papers published by ACM in four domains of Computer Science, Distributed Systems, Information Search and Retrieval, Learning, and Social and Behavioral Sciences, and it consists of a total of 10,935 documents. This dataset has only been used to create a labeling corpus (see Section 3.2).

4.2. Experimental Setup. The following evaluation metrics will be employed in the experiments:

(i) Precision is the fraction of relevant instances among the retrieved instances.

$$\text{Precision} = \frac{tp}{tp + fp}, \tag{2}$$

where tp and fp denote true positives and false positives, respectively.

(ii) Recall is the fraction of relevant instances that were retrieved.

$$\text{Recall} = \frac{tp}{tp + fn}, \tag{3}$$

where fn denotes false negatives.

(iii) F – measure is the harmonic mean of Precision and Recall.

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

5. Experimental Results and Discussion

We follow experiments in the same sequence as mentioned in Section 3.

5.1. POS Tagging. As discussed in Section 3.1, Table 2 presents different combinations of POS experimented for the task of keyword extraction.

Here, NO = Nouns, AD = Adjectives, F = Foreign Words, I = Irrelevant Terms, V = Verbs, NU = Numbers, G = Genitive Markers, AG = Agents, and MV = Modal Verbs.

Foreign Words include non-English words, and Irrelevant Terms are represented by a union of all prepositions, conjunctions, determiners, possessive pronouns, particles, adverbs, and interjections [45], while Genitive Markers show ownership, measurement, association, or source, e.g., "boy's" and "of the boy."

Although the combination of POS selected for our methodology ranks fourth among the different ones experimented, we, for obvious reasons, avoided those combinations that included either foreign words or irrelevant terms.

5.2. Labeling Corpus. As discussed in Section 3.2, Table 3 displays various frequencies of terms assigned as labels in the labeling corpus, which were experimented for the task of keyword extraction.

5.3. Ratio Metric. As discussed in Section 3.3, we experimented with different threshold values for x for the task of keyword extraction, and found $x = 5$ to be the optimal value in terms of F-measure, as mentioned in Table 4.

All results related to different stages for the process of keyword extraction are summarized in Table 5, as discussed in Sections 3.1–3.3.

Here, $F1$, $F2$, and $F3$ represent the POS Tagging, Labeling Corpus, and Ratio Metric features, respectively.

5.4. Keyphrase Extraction. As discussed in Section 3, the optimal values yielded for the first three features for the process of keyword extraction were then revisited to yield the optimal values for the process of keyphrase extraction. Although the same optimal values were yielded for the first two features, the Ratio Metric feature produced an optimal value at $x = 8$, as mentioned in Table 6, and also reflected in Tables 4 and 5.

This is the simplest combination of keyword extraction and keyphrase extraction processes where all adjacently located keywords in d were utilized to form keyphrases.

As discussed in Section 3.4, for our second combination of keyword extraction and keyphrase extraction processes, we experimented with different values for P, and found $P = 55$ to be the optimal value in terms of F-measure, as mentioned in Table 7.

Table 3: Various frequencies of terms as labels in Labeling Corpus.

S#		Frequency of Term as Label	Precision	Recall	F – measure
1		**$f \geq 1$**	**0.4106**	**0.8194**	**0.5471**
2		$1 \leq f \leq 15$	0.3277	0.3237	0.3257
3	Top-N, N = All	$1 \leq f \leq 28$	0.3504	0.4363	0.3886
4		$1 \leq f \leq 80$	0.3821	0.6466	0.4803
5		$1 \leq f \leq 294$	0.4030	0.7796	0.5313

Bold values represent the best results achieved in terms of F-measure.

Table 4: Different threshold values for x.

S#		x	Precision	Recall	F – measure
1		**5**	**0.6115**	**0.7195**	**0.6611**
2	Top-N, N = All	6	0.5901	0.7497	0.6604
3		7	0.5716	0.7704	0.6563
4		8	0.5589	0.7833	0.6523

Bold values represent the best results achieved in terms of F-measure.

Table 5: Different stages for the process of keyword extraction.

S#		Stage	Precision	Recall	F – measure	Change in F – measure
1		CSS	0.2790	0.9073	0.4268	—
2	Top-N, N = All	CSS + $F1$	0.3303	0.8195	0.4708	↑4%
3		CSS + $F1$ + $F2$	0.4106	0.8194	0.5471	↑8%
4		**CSS + F1 + F2 + F3**	**0.5589**	**0.7833**	**0.6523**	**↑11%**

Bold values represent the best results achieved in terms of F-measure.

Table 6: Different threshold values for x.

S#		x	Precision	Recall	F – measure
1		4	0.2174	0.4441	0.2919
2		5	0.2269	0.5156	0.3152
3	Top-N, N = All	6	0.2274	0.5491	0.3217
4		7	0.2274	0.5775	0.3263
5		8	0.2276	0.5953	0.3293
6		9	0.2249	0.6064	0.3282

Table 7: Different values for P.

S#	$P1$	Precision	Recall	F – measure
1	40	0.4316	0.4730	0.4513
2	50	0.4031	0.5597	0.4687
3	**55**	**0.3927**	**0.5820**	**0.4690**
4	60	0.3764	0.6028	0.4635
5	70	0.3418	0.6366	0.4447

Bold values represent the best results achieved in terms of F-measure.

As discussed in Section 3.4, for our third combination of keyword extraction and keyphrase extraction processes, we experimented with different values for $P1$, and found $P1 = 59$ to be the optimal value in terms of F-measure, as mentioned in Table 8.

As discussed in Section 3.4, for our fourth combination of keyword extraction and keyphrase extraction processes, we experimented with different values for $P2$, and found $P2 = 55$ to be the optimal value in terms of F-measure, as mentioned in Table 9.

Table 8: Different values for $P1$.

S#	$P1$	Precision	Recall	F – measure
1	40	0.4048	0.4469	0.4248
2	50	0.3965	0.5556	0.4628
3	59	0.3818	0.6075	0.4689
4	60	0.3791	0.6138	0.4687
5	70	0.3502	0.6592	0.4574

Table 9: Different values for $P2$.

S#	$P2$	Precision	Recall	F – measure
1	40	0.4294	0.4746	0.4509
2	50	0.4075	0.5700	0.4752
3	55	0.3984	0.5954	0.4774
4	60	0.3837	0.6207	0.4742
5	70	0.3516	0.6616	0.4592

All results related to different combinations of keyword extraction and keyphrase extraction processes are summarized in Table 10, as discussed in Section 3.4.

Table 10: Different combinations of keyword extraction and keyphrase extraction processes.

S#	Combination of processes	Precision	Recall	$F-$measure	Change in $F-$measure
1	Keyword extraction + adjacent keywords in d to form keyphrases	0.2276	0.5953	0.3293	—
2	Keyword extraction + adjacent keywords in d to form top-P% keyphrases	0.3927	0.5820	0.4690	↑13%
3	Keyword extraction (top-P1% unigrams) + adjacent keywords in d to form keyphrases	0.3818	0.6075	0.4689	—
4	**Keyword extraction (top- P1% unigrams) + adjacent keywords in d to form top-P2% keyphrases**	**0.3984**	**0.5954**	**0.4774**	**↑1%**

Bold values represent the best results achieved in terms of F-measure.

Table 11: Systematic comparison of CCSS with other techniques over INSPEC dataset.

S#	Technique	Precision	Recall	$F-$measure
1	PageRank [18] evaluated by Liu et al. [23]	0.3300	0.1710	0.2250
2	Term frequency inverse document frequency (TFIDF) evaluated by Liu et al. [23]	0.3330	0.1730	0.2270
3	LDA [22] evaluated by Liu et al. [23]	0.3320	0.1720	0.2270
4	TPR [23]	0.3540	0.1830	0.2420
5	Liu et al. [24]	0.3270	0.6530	0.4360
6	SemanticRank [25]	0.4210	0.5320	0.4700
7	CCSS	0.3984	0.5954	0.4774

5.5. CCSS Vs. State-of-the-Art Techniques. We systematically compared the performance of CCSS with other techniques, when implemented over INSPEC dataset, as presented in the literature, and such analysis is presented in Table 11. It is clear that CCSS has outperformed all state-of-the-art keyphrase extraction techniques presented in the literature.

6. Conclusion and Future Work

In this paper, we have presented a novel supervised technique for extraction of keywords from medium-sized documents, namely Corpus-based Contextual Semantic Smoothing (CCSS). CCSS extended the concept of Contextual Semantic Smoothing (CSS), which considered term usage patterns in similar texts to improve term relevance information. We introduced four more features beyond CSS as our novel contributions in this work. We systematically compared the performance of CCSS with other techniques, when implemented over INSPEC dataset, where CCSS clearly outperformed all state-of-the-art keyphrase extraction techniques presented in the literature.

Our future work includes utilizing CCSS in the applications of indexing and search, summarization, and multilingual summarization, of medium-sized documents. We are also currently working on compiling the literature review for all keyword extraction-based applications beyond and including the abovementioned ones.

Data Availability

Previously reported INSPEC and ACM datasets were used to support this study and are available at https://www.theiet.org/publishing/inspec/and https://www.innovalor.nl/, respectively. The datasets used in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors gratefully acknowledge both Mohammad Ali Jinnah University (MAJU), Karachi, Pakistan, and Deanship of Research, Islamic University of Madinah, Madinah, Kingdom of Saudi Arabia, on the support provided for this research. This research was funded by Deanship of Research, Islamic University of Madinah, Madinah, Kingdom of Saudi Arabia.

References

[1] V. Qazvinian, D. R. Radev, and A. Özgür, "Citation summarization through keyphrase extraction," in *Proceedings of the 23rd International Conference on Computational Linguistics (Beijing, China) (COLING '10)*, pp. 895–903, Association for Computational Linguistics, Stroudsburg, PA, USA, August 2010.

[2] H. Zha, "Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Tampere, Finland) (SIGIR '02)*, pp. 113–120, ACM, New York, NY, USA, August 2002.

[3] A. Hliaoutakis, K. Zervanou, G. Euripides, E. G. M. Petrakis, and E. E. Milios, "Automatic document indexing in large medical collections," in *Proceedings of the International Workshop on Healthcare Information and Knowledge Management (Arlington, VA, USA) (HIKM '06)*, pp. 1–8, ACM, New York, NY, USA, November 2006.

[4] Z. Bar-Yossef and M. Gurevich, "Estimating the impressionrank of web pages," in *Proceedings of the 18th International Conference on World Wide Web (Madrid, Spain)*,

WWW '09). ACM, pp. 41–50, New York, NY, USA, April 2009.

[5] O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 – Volume 3 (Singapore) (EMNLP '09)*, pp. 1318–1327, Association for Computational Linguistics, Stroudsburg, PA, USA, August 2009.

[6] W. Wu, B. Zhang, and M. Ostendorf, "Automatic generation of personalized annotation tags for twitter users," in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Los Angeles, CA, USA) (HLT '10)*, pp. 689–692, Association for Computational Linguistics, Stroudsburg, PA, USA, June 2010.

[7] K. S. Dave and V. Varma, "Pattern based keyword extraction for contextual advertising," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (Toronto, ON, Canada) (CIKM '10)*, pp. 1885–1888, ACM, New York, NY, USA, October 2010.

[8] W.-T. Yih, J. Goodman, and V. R. Carvalho, "Finding advertising keywords on web pages," in *Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland) (WWW '06)*, pp. 213–222, ACM, New York, NY, USA, May 2006.

[9] C. Wartena, R. Brussee, and W. Slakhorst, "Keyword extraction using word co-occurrence," in *Proceedings of the 2010 Workshops on Database and Expert Systems Applications (DEXA '10) (Bilbao, Spain)*, pp. 54–58, IEEE Computer Society, Washington, DC VA, USA, August 2010.

[10] O. A. Khan and A. Karim, "MIKE: an interactive microblogging keyword extractor using contextual semantic smoothing," in *Proceedings of the 24th International Conference on Computational Linguistics (Mumbai, India) COLING '12*, pp. 289–296, Association for Computational Linguistics, Stroudsburg, PA, USA, December 2012.

[11] X. Wu, F. Xie, G. Wu, and W. Ding, "Personalized news filtering and summarization on the web," in *Proceedings of the 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence (ICTAI '11)*, pp. 414–421, Boca Raton, FL, USA, November 2011.

[12] S. A. Hossain, A. S. M. M. Rahman, T. T. Tran, and A. E. Saddik, "Location aware question answering based product searching in mobile handheld devices," in *Proceedings of the 2010 IEEE/ACM 14th International Symposium on Distributed Simulation and Real Time Applications (DS-RT '10)*, pp. 189–195, Fairfax, VA, USA, October 2010.

[13] K. Zoltán and S. Johann, "Semantic analysis of microposts for efficient people to people interactions," in *Proceedings of the 2011 RoEduNet International Conference 10th Edition: Networking in Education and Research*, pp. 1–4, Iasi, Romania, June 2011.

[14] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning, "Domain-specific keyphrase extraction," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence, (IJCAI '99)*, pp. 668–673, San Francisco, CA, USA, July 1999.

[15] F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," in *Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '09) (Boulder, CO)*, pp. 620–628, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009.

[16] F. Liu, F. Liu, and Y. Liu, "A supervised framework for keyword extraction from meeting transcripts," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 3, pp. 538–548, March 2011.

[17] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP '03) (Sapporo, Japan)*, pp. 216–223, Association for Computational Linguistics, Stroudsburg, PA, USA, July 2003.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd, *"The PageRank Citation Ranking: Bringing Order to the Web"*, Stanford InfoLab, Stanford, CA, USA, 1999.

[19] R. Mihalcea and P. Tarau, "Textrank: bringing order into texts," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*, pp. 404–411, (Barcelona, Spain), Association for Computational Linguistics, Stroudsburg, PA, USA, July 2004.

[20] X. Wan and J. Xiao, "Collabrank: towards a collaborative approach to single-document keyphrase extraction," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)*, (Manchester UK), pp. 969–976, Association for Computational Linguistics, Stroudsburg, PA, USA, August 2008.

[21] X. Wan and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge," in *Proceedings of the 23rd National Conference on Artificial Intelligence: Volume 2 (AAAI '08)*, pp. 855–860, Chicago, IL, USA, July 2008.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, January 2003.

[23] Z. Liu, W. Huang, Y. Zheng, and M. Sun, "Automatic keyphrase extraction via topic decomposition," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10) (Cambridge, MA, USA)*, pp. 366–376, Association for Computational Linguistics, Stroudsburg, PA, USA, October 2010.

[24] Z. Liu, P. Li, Y. Zheng, and M. Sun, "Clustering to find exemplar terms for keyphrase extraction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 – Volume 1 (EMNLP '09) (Singapore)*, pp. 257–266, Association for Computational Linguistics, Stroudsburg, PA, USA, August 2009.

[25] G. Tsatsaronis, I. Varlamis, and K. Nørvåg, "SemanticRank: ranking keywords and sentences using semantic graphs," in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1074–1082, Association for Computational Linguistics, Stroudsburg, PA, USA, August 2010.

[26] L. Zhang, J. Wu, Y. Zhuang, Y. Zhang, and C. Yang, "Review-oriented metadata enrichment: a case study," in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '09)*, pp. 173–182, ACM, New York, NY, USA, June 2009.

[27] M. Wenchao, L. Lianchen, and D. Ting, "A modified approach to keyword extraction based on word-similarity," in *Proceedings of the 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICICISYS '09)*, pp. 388–392, Shanghai, China, November 2009.

[28] A. Panunzi, M. Fabbri, and M. Moneglia, "Keyword extraction in open-domain multilingual textual resources," in *Proceedings of the First International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution AXMEDIS'05*, p. 4, Florence, Italy, November 2005.

[29] T. Takehara, S. Miki, N. Nitta, and N. Babaguchi, "Extracting Context Information from Microblog Based on Analysis of Online Reviews," in *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo (ICME '12) Workshops*, pp. 248–253, Melbourne, VIC, Australia, July 2012.

[30] J. Tang, Z. Liu, M. Sun, and J. Liu, "Portraying user life status from microblogging posts," *Tsinghua Science and Technology*, vol. 18, no. 2, pp. 182–195, April 2013.

[31] J. Liu and J. Wang, "Keyword extraction using language network," in *Proceedings of the 2007 International Conference on Natural Language Processing and Knowledge Engineering (NLPKE '07)*, pp. 129–134, Beijing, China, August 2007.

[32] Y. Ouyang, W. Li, and R. Zhang, "273. Task 5. Keyphrase extraction based on core word identification and word expansion," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10). Association for Computational Linguistics*, pp. 142–145, Stroudsburg, PA, USA, July 2010.

[33] J. Hauffa, T. Lichtenberg, and G. Groh, "Towards an NLP-based topic characterization of social relations," in *Proceedings of the 2012 International Conference on Social Informatics (ICSI '12),*, pp. 289–294, Alexandria, VA, USA, December 2012.

[34] G. Berend and R. Farkas, "SZTERGAK: feature engineering for keyphrase extraction," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, pp. 186–189, Association for Computational Linguistics, Stroudsburg, PA, USA, July 2010.

[35] Y. J. Lui, R. Brent, and A. Calinescu, "Extracting Significant Phrases from Text," in *Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, pp. 361–366, Niagara Falls, ON, Canada, May 2007.

[36] G. Fang, C. Yuan, X. Wang, J. Li, and Z. Song, "From keywords to social tags: tagging for dialogues," in *Proceedings of the 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE '11)*, pp. 106–113, Tokushima Japan, November 2011.

[37] C. Chitra and A. Julian, "Searching video blogs with integration of context and content analysis," in *Proceedings of the 2010 International Conference on Innovative Computing Technologies (ICICT) '10)*, pp. 1–5, Karur, India, February 2010.

[38] G. B. Colombo, M. J. Chorley, V. Tanasescu, S. M. Allen, C. B. Jones, and R. M. Whitaker, "Will you like this place? A tag-based place representation approach," in *Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pp. 224–229, San Diego, CA, USA, March 2013.

[39] A. Hulth, "Enhancing linguistically oriented automatic keyword extraction," in *Proceedings of the Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Short Papers (HLT-NAACL-Short '04)*, pp. 17–20, Association for Computational Linguistics, Stroudsburg, PA, USA, May 2004.

[40] L.-F. Chien, "PAT-tree-based keyword extraction for Chinese information retrieval," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, pp. 50–58, ACM, New York, NY, USA, July 1997.

[41] E. Pianta and S. Tonelli, "KX: a flexible system for keyphrase extraction," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, pp. 170–173, Association for Computational Linguistics, Stroudsburg, PA, USA, July 2010.

[42] P. Treeratpituk, P Teregowda, J. Huang, and C. Lee Giles, "SEERLAB: a system for extracting key phrases from scholarly documents," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, pp. 182–185, Association for Computational Linguistics, Stroudsburg, PA, USA, July 2010.

[43] S. Muresan, E. Tzoukermann, and J. L. Klavans, "Combining linguistic and machine learning techniques for email summarization," in *Proceedings of the 2001 Workshop on Computational Natural Language Learning - Volume 7 (ConLL '01)*, p. 8, Stroudsburg, PA, USA, July 2001.

[44] M. Paukkeri, A. P. García-Plaza, S. Pessala, and T. Honkela, "Learning taxonomic relations from a set of text documents," in *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT '10)*, pp. 105–112, Wisla, Poland, October 2010.

[45] K. M. Carley, "AutoMap User's Guide," *Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International (ISRI), School of Computer Science, Carnegie Mellon University*, June 2013.