International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.5 (2015), pp.355-362
http://dx.doi.org/10.14257/ijmue.2015.10.5.33

Exploiting YouTube's Video ASR Scripts to Extend Educational Videos Textual Representative Tags Based on Gibb's Sampling Technique

Ambele Robert Mtafya[1], Dongjun Huang[2] and Gaudence Uwamahoro[3]

[1]*Central South University, Changsha, Hunan, China;*
Dar es salaam Institute of Technology, Tanzania
[2,3]*Central South University, Changsha, Hunan, China*
[1]*kakaambe@gmail.com*

Abstract

Given the importance of the textual information in content retrieval, it is desirable that the textual representation of educational videos contents in social media platforms like YouTube capture the semantics of what is really in content they represent. Such coherent textual representations are important in objective video content retrieval, repurposing, reuse and sense- making of the content. In this study,the Automatic Speech Recognition (ASR) in the video tracks was leveraged to supplement the insufficient video content representations done through video title alone. The Latent Dirichlet allocation (LDA) implementation of Gibb's sampling topic modeling approach was used to evaluate the suitability of various textual representations for YouTube educational videos and extract the candidate topic that extends well the original YouTube keywords. The results show that in topics space, YouTube ASR script performs well as a representative textual source in dominant topic than the combined textual representations. The automatic keywords extension obtained using our method add value to applications that use tags for content discovery or retrieval

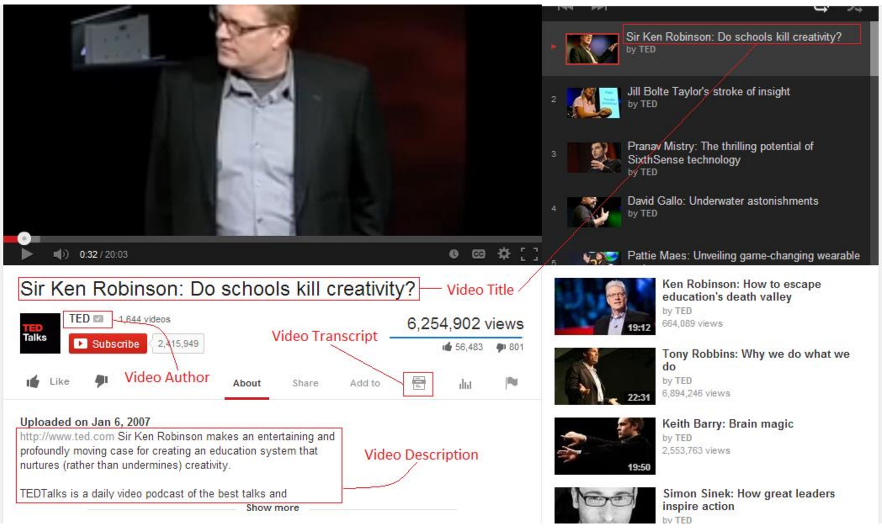
Keywords: *content discovery; textual representation; Gibb's sampling; Video ASR scripts; topic modeling*

1. Introduction

The most widespread way in which video information is retrieved is through use combination of video metadata (title, author, date, duration, format, etc.) and user-generated description (user tags, ratings, reviews, etc.) [1]. This approach is also the basis for navigation through video archives in systems such as the famous video content sharing platform YouTube, Open Video project and the Internet Archives. However for objective video content retrieval, repurposing, reuse and sense-making of the massive data in online social platforms like YouTube, there is a need of exploiting the content-based semantics as well as the available embedded social dimension objects like user comments, video rating, and video like and dislike metrics. As it can be seen from figure 1, a typical YouTube video has; users' comments; likes or unlike; video's views; time watched and subscription driven statistics also text transcript for some videos. The author supplied tags describing the content though are used in content retrieval, are usually hidden from the user; they can only be extracted via an API. The YouTube registered users can only contribute through commenting or rating of the video content via the like button. The absence of user driven tags and the noisy nature of the user comments makes verbose search to rely on metadata only. This poses a major drawback in resource discovery applications because video metadata textual presentation alone is not good enough

ISSN: 1975-0080 IJMUE

representative of the video content; hence motivating further research on exploiting audio-visual features[2].


0:32 / 20:03
Sir Ken Robinson: Do schools kill creativity?
Video Title
TED
1,644 videos
TED Talks
Subscribe
2,415,949
6,254,902 views
Video Transcript
56,483
801
Like
Video Author
About
Share
Add to
Show more
Video Description
Sir Ken Robinson: Do schools kill creativity?
by TED
2
Jill Bolte Taylor's stroke of insight
by TED
3
Pranav Mistry: The thrilling potential of SixthSense technology
by TED
4
David Gallo: Underwater astonishments
by TED
5
Pattie Maes: Unveiling game-changing wearable
Ken Robinson: How to escape education's death valley
by TED
664,089 views
19:12
Tony Robbins: Why we do what we do
by TED
6,894,246 views
22:31
Keith Barry: Brain magic
by TED
2,553,763 views
19:50
Simon Sinek: How great leaders inspire action
by TED

Uploaded on Jan 6, 2007

http://www.ted.com Sir Ken Robinson makes an entertaining and profoundly moving case for creating an education system that nurtures (rather than undermines) creativity.

TEDTalks is a daily video podcast of the best talks and

Figure 1. Metadata from YouTube video

Considering the strength of audio features, YouTube in 2009 introduced the automatic closed captioning service for videos not longer than 20 minutes and whose languages are supported[1]. The introduced captions technology was envisioned to not only helps the deaf and hearing impaired, but also to boosts content access, navigation and machine translation as well. Google combined its automatic speech recognition (ASR) technology with the YouTube caption system to offer automatic captions which use the same voice recognition algorithms as Google Voice to automatically generate captions for video. However, the ASR methods are known for poor performance [3], especially if the speaker accent is not as the one used in training the language model in use.

Nevertheless, studies on information retrieval based on ASR text suggested that reasonable information retrieval can achieved based on the ASR text; for example Désilets *et al.* (2000) produced accurate key-phrases for transcriptions with Word Error Rates (WER) of the order of 25% and according to Hauptmann and Wactlar study [4], word error rates up to 25% did not significantly impact information retrieval and error rates of 50% still provided 85–95% of the recall and precision relative to fully accurate transcripts in the same retrieval system. Furthermore Hank Liao *et al.* [5], describes recent improvements to the original YouTube ASR system, in particular the use of owner-uploaded video transcripts to generate additional semi-supervised training data and deep neural networks acoustic models with large state inventories in which they had an improved performance by about 13% as compared to previously reported sequence trained DNN results for this task. This is a motivation that the automatic annotation of the UGC can safely be done with the existing YouTube ASR system; the only major challenge remaining is getting the significant textual representation of the content. In this research we first evaluate implicitly the goodness of the YouTube ASR script as the textual source for the content it represents and use the best representation as the basis for extending the hidden official YouTube tags.

2. Related Work

Most of the ASR-based keyphrases extraction studies like [6-8] hinges on semantic relatedness mainly exploiting the reference semantics from the web content repository

[1] http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html

like WorldNet and Wikipedia. Chen *et al.* [9] describes general semantic relatedness approaches on keyphrases extraction area, though of recent, the trend is towards use of LDA [10, 11]. On minimizing ASR transcription errors, Diana Inkpen *et al.* [12] used PMI scores as semantic measure to filter out the mis-transcribed words from keyphrases extraction using a subset of the ABC and PRI stories of the TDT2 English audio data that had correct transcripts generated by humans. It can be noted that not many studies on keyphrases extraction area had used ASR scripts as their textual source; we fill that gap by evaluating the goodness of the YouTube educational videos textual content representations and exploiting them for automatic tagging. Our approach aims at broadcasting the relevant representative semantics leveraging both the metadata and the automatic speech recognition (ASR) in the video tracks. One advantage of text-based approaches is that they can utilize the well proved methods for text document analysis [13] including the popular LDA that implements the topic modeling based on Gibbs sampling [14].For completeness, the LDA as described in [15] is summarized here. The main idea for topic modeling is to use the observed documents to infer the hidden topic structure. In LDA the observed variables are the words of the documents; the hidden variables are the topic structure. The generative process for LDA corresponds to the following joint distribution of the hidden and observed variables, defined in equation 1 as:

$$p(\beta,\theta,z,w) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d)\left(\prod_{n=1}^{N} p(z_{d,n}\,|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})\right) \quad (1)$$

Where

$\beta_{1:K}$, are the topics.

θ_d, are the topic proportions for the d^{th} document,

z_d , are the topic assignments for the d^{th} document and

w_d, are the observed words for document d

The central computational problem lies in computing the conditional distribution of the topic structure given the observed documents:

$$p(\beta,\theta,z|w) = \frac{p(\beta,\theta,z,w)}{p(w)} \qquad (2)$$

Due to exponential size of the possible topic number of structures, equation 2 is solved approximately either by variation methods or sampling. The most commonly used sampling algorithm for topic modeling is Gibbs sampling; Gibbs sampling procedure considers each word token in the text collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on the topic assignments to all other word tokens. From this conditional distribution, a topic is sampled and stored as the new topic assignment for this word token[16].

3. Method Description

The TED's 2013 list of twenty most viewed talks were chosen for the study since they are good representative of the educational content we envision to generalize; in each video a single speaker is identified in the presentation. The other advantages of using TED videos lies in the fact that they are firstly accompanied with official script and secondly the same video content in duplicated in their YouTube channel which makes evaluative comparison easy .The official scripts were taken from the corresponding videos at TED's website while the automatically generated scripts were extracted from the corresponding posting on their YouTube channel. The conceptual frame work of our study is given in Figure 2.

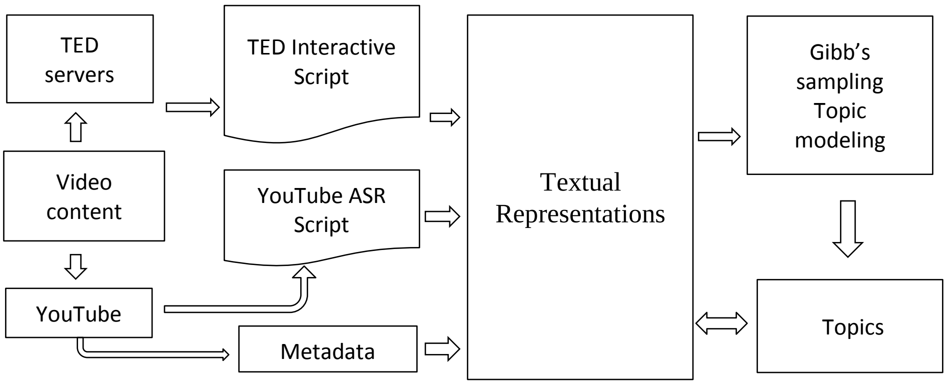


Figure 2. Conceptual framework

For each video source, its corresponding video on TED network and YouTube were consulted to extract the official script and the ASR script respectively. In addition, the video metadata as described in figure 1 were extracted together with the hidden official keywords using a Python based API. A novel keyphrasing algorithm[17] was used on the three textual sources to extend the number of textual representations for each video content to nine; namely the three originals, their three keyphrases representation and the official keyphrases from YouTube resource. The extra two sources were obtained by taking unions of the ASR keywords with metadata and the official keywords for comparison purposes. A sample top20 keywords representation for the first video titled "Sir Robinson: Does education kill creativity?" is given in Table 1.

Table1. Sample Keywords Textual Representations

YouTube keywords	*Metadata keywords*	*TED keywords*	*ASR keywords*
'Ken', 'Robinson', 'TED', 'TEDTalks', 'Talks', 'TED2006', 'education', 'educational', 'system', 'creativity', 'innovation', 'schooling', 'school', 'curiosity'	*' al gore ', 'talk ', 'best talks ', 'tedtalks ', 'tedtalks cover ', 'com', 'nurtures ', 'performances ', 'ted conference', 'leading thinkers 'doers ', 'lives ', ' hans rosling' 'arthur benjamin' ' ted stands ', 'technology', ' entertainment', 'design'*	*'education system ', 'dance school', 'human creativity , 'human brain', ' children dance ', 'human ecology', 'human capacity', 'children grow ', 'william shakespeare', 'royal ballet ', 'drawing lesson', 'blood run ' 'stigmatize mistakes', 'man speaks ', 'visit education', 'think math ', 'human imagination', 'university professors'*	*'education system ', 'dance school', 'face think ', 'think matt ', 'fact creativity ', 'life sarah ', 'party joseph ', 'status thank ', 'see comes ', 'subjects everyone ', 'stigmatize mistakes', 'man speaks ', 'body experiences ', 'son watcher', 'human creativity', 'human capacity', 'human brain ', 'life affection*

4. Experimental Result

4.1 Textual Representations

The textual sources were then used in MAchine Learning for LanguagE Toolkit (MALLET)[2] which implements Latent Dirichlet Allocation based on Gibb's sampling. Experts in the topic modeling believe that sampling based LDA is a more accurate fitting method than the variational Bayes [18] which are easier to parallelize and guaranteed to converge but they essentially solve an approximate problem [19]. Each set of textual representation of the given video were mapped into a space with twenty topics; the normalized topical proportion contributions were analyzed for the dominant topic the result is given in Table 2. The dominant topics like the one in Figure 3 were tested as queries' in Google video search to prove that the topic indeed represent the corresponding content; all returned the relevant results. Visual comparison of YouTube ASR with the corresponding official TED script and other textual representations are shown in Figures 4-6.

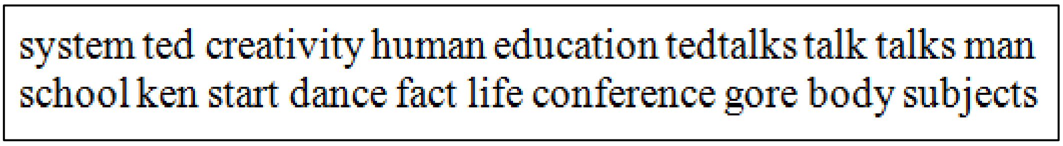

system ted creativity human education tedtalks talk talks man school ken start dance fact life conference gore body subjects

Figure 3. Sample Dominant Topic for the Top Video

Table 2. The Normalized Topical Proportion Contributions

Textual Representation	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
Official TED	0.061	0.487	0.217	0.006	0.376	0.508	0.426	0.224	0.130	0.134
ASR script	0.133	0.469	0.202	0.022	0.304	0.434	0.335	0.193	0.140	0.118
Keywords-ASR	0.878	1.000	0.887	0.676	0.804	1.000	0.538	0.727	0.898	0.910
keywords-ASR + Metadata	0.922	0.909	0.469	0.811	0.559	0.795	0.339	0.579	0.519	0.584
Keywords-ASR+Metadata +YTkywds	0.923	0.854	0.346	0.886	0.425	0.056	0.413	0.455	0.410	0.635
keywords-Official TED	0.483	0.960	0.173	0.141	0.854	0.893	0.826	0.706	0.912	0.618
Metadata	0.805	0.750	0.010	0.395	0.205	0.406	0.128	0.107	0.017	0.009
keywords-Youtube Metadata	0.827	0.000	0.032	0.738	0.133	0.500	0.104	0.000	0.035	0.028
Keywords-Youtube	0.749	0.000	0.110	0.582	0.195	0.000	0.484	0.000	0.197	0.707

From the score analysis of dominant topic proportions for each textual representation, the high pattern correlation between the ASR script and the official TED can be noted (Figure 4). The pattern signify that the two textual representation of the video content are very close near-

[2] http://mallet.cs.umass.edu/

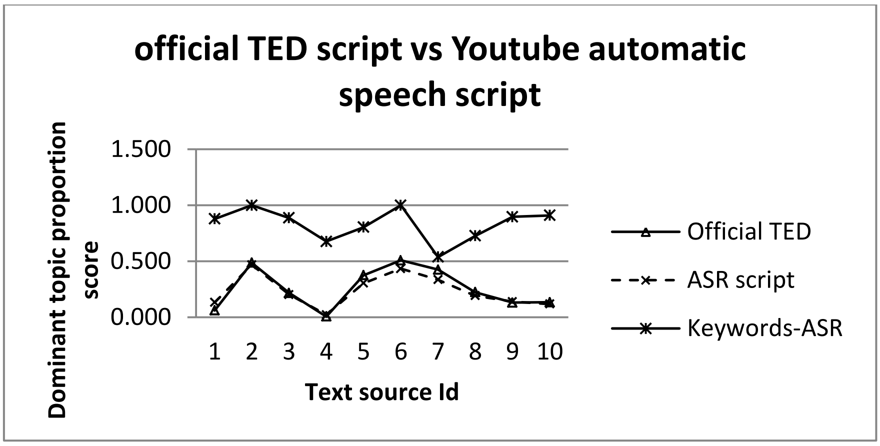


Figure 4. Comparison between ASR Script and Official TED Script

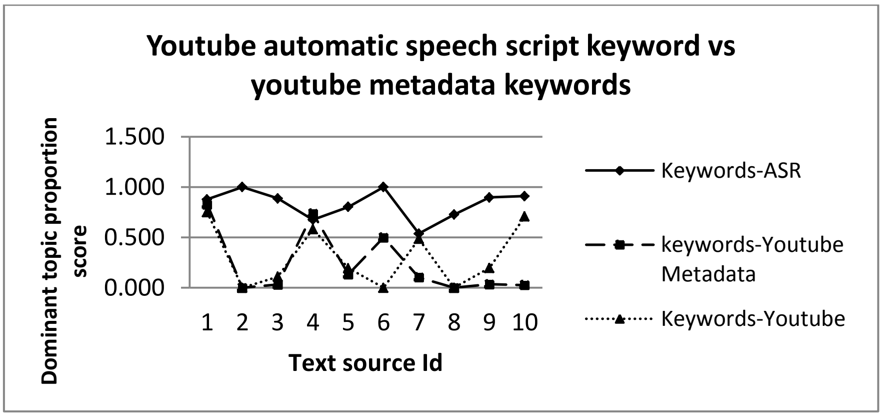


Figure 5. Comparison between ASR Derived Presentation and YouTube Native Text Presentation

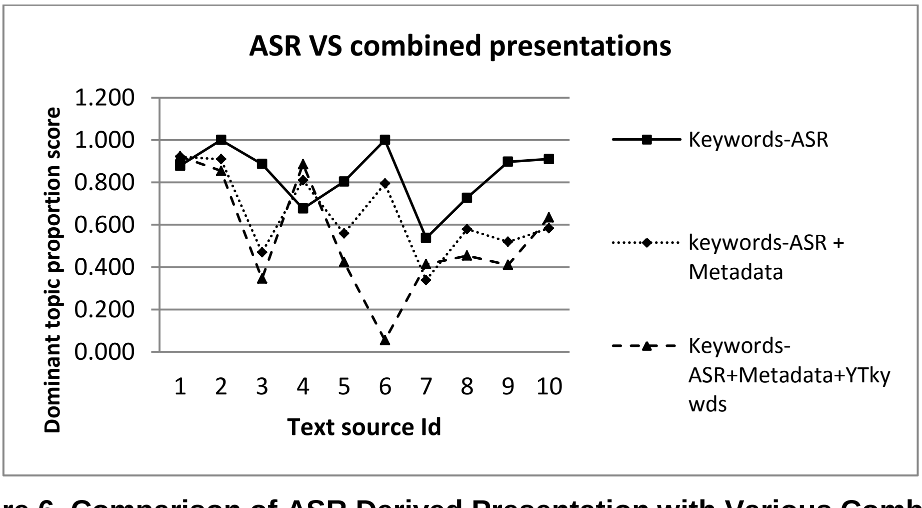


Figure 6. Comparison of ASR Derived Presentation with Various Combined Presentations

duplicate documents, therefore the minor differences arising from YouTube ASR mis-transcription errors can be ignored in topics' space. It can further be noted that both the ASR script and the official TED textual representation in figure 4 scores less as compared to the ASR keywords representation leading to a conclusion that ASR keywords representation is the better one among the three. The same ASR keywords textual representation proportional contribution to the dominant topic is seen to be superior to YouTube's native metadata and keywords (Figure 5); the observation validates the hypothesis that native YouTube metadata are sparse and do not represent well the actual video content and adds significance of this study. On the other hand it is surprising to see that the combined textual representations scores low as compared to ASR even the union representation that has three components, including the ASR keywords in it (Figure 6).

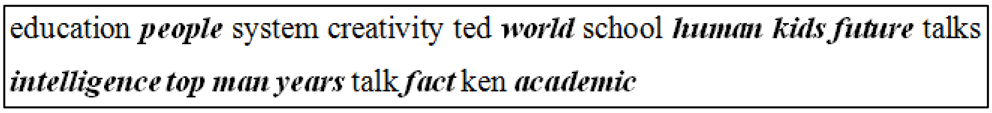
4.2 The Algorithm for Extending the YouTube Keywords

The experimental results in section 4.1 motivate the use of textual representations entirely from YouTube. For automatic YouTube's keywords extension in topics space, the algorithm that leverage the YouTube's ASR text and it's metadata as the key textual representation is presented together with the sample result of its implementation:

The YoutubeKeywordsExtension algorithm

```
    INPUT: YouTube video Id
    OUTPUT: Extended keywords for the input Video ID
    Let V= Set of YouTube Ids of instructional content (where a single speaker is
recognized)
    BEGIN
1.  READ  V  FROM source
2.  SET  [ ] ← extended_keywds
3.  FOR each v ϵ V  DO
a.  READ   auto, meta, yt_keywds  FROM  source
      // YouTube's automatic script, metadata and YouTube keywords
b.  CALL  KeywordsExtractor (auto)  // as discussed in[17]
            RETURN auto_keywords
c.  CALL Mallet (auto, auto_keywd, meta)
            RETURN {topic_keys}, {topics composition}
d.  dominant_ topic = max{topics composition}
e.  IF topic_key  IN  dominant_ topic
            THEN
               RETURN dominant_topic_text
            ENDIF
f.  keywds ←∪ {yt_keywds, dominant_topic_text}
g.  extended_keywds ← extended_keywds.append((v, keywds))
            ENDFOR
      END
```

When this algorithm was implemented on the first video on TED's 2013 top 20 shows list the resulting dominant topic returned is as seen in figure 7; the bolded words in the figure identifies words that are in this topic but not in the original YouTube assigned keywords. Namely the keywords extension candidate words are: *people, world, human, kids, future, intelligence, top, man, years, fact* and *academic*.



education ***people*** system creativity ted ***world*** school ***human kids future*** talks ***intelligence top man years*** talk ***fact*** ken ***academic***

Figure 7. Candidate Words for Keywords Extension (in Bold)

5. Conclusion

In this study, the LDA implementation of Gibb’s sampling topic modeling approach was used to evaluate the suitability of various textual representations for YouTube educational videos and extract the candidate topic that extends well the original YouTube keywords. Without loss of generality it can be concluded that; when measured in sampled topics space, the YouTube ASR keywords proved to be a better textual representation as compared to other textual representations considered. When this better textual representation was used in Gibbs’s sampling implementation in Mallet in conjunction with the YouTube’s video metadata, the resulting dominant topic contained very interesting extension to the YouTube keywords. Further study will involve exploiting YouTube script sampled topics for content navigation and recommendation.

References

[1] A. F. Smeaton, "Information Systems", vol. 32, no. 4, (**2007**). p. 545-559.

[2] C. Eickhoff, W. Li, and A. P. de Vries, "Exploiting User Comments for Audio-Visual Content Indexing and Retrieval", in Advances in Information Retrieval Springer, (**2013**), pp. 38-49.

[3] A. G. Hauptmann, R. Jin, and T. D. Ng, "Multi-Modal Information Retrieval from Broadcast Video Using Ocr and Speech Recognition", in JCDL'02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, (**2002**), pp. 160-161.

[4] A. G. Hauptmann and H. D. Wactlar, "Indexing and Search of Multimodal Information", in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, 21-24 April 1997, (**1997**), Los Alamitos, CA, USA.

[5] H. Liao, E. McDermott, and A.W. Senior, Large Scale Deep Neural Network Acoustic Modeling with Semi-Supervised Training Data for Youtube Video Transcription, in ASRU (**2013**), IEEE. p. 368-373.

[6] H. Feng, Z. Pang, K. Qiu, and G. Song, "Web-Based Semantic Analysis of Chinese News Video", in Advances in Multimedia Information Processing - Pcm 2006, Y. Zhuang, S.-Q. Yang, Y. Rui, and Q. He, Editors Springer Berlin Heidelberg (**2006**). p. 502-509.

[7] S. Cucerzan. "Large-Scale Named Entity Disambiguation Based on Wikipedia Data". in EMNLP-CoNLL. (**2007**).

[8] O. Medelyan and I. H. Witten, "Thesaurus Based Automatic Keyphrase Indexing", in Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (**2006**).

[9] Y.-N. Chen, Y. Huang, S.-Y. Kong, and L.-S. Lee, "Automatic Key Term Extraction from Spoken Course Lectures Using Branching Entropy and Prosodic/Semantic Features", in Spoken Language Technology Workshop (SLT), 2010 IEEE, (**2010**).

[10] F. El-Ghannam and T. El-Shishtawy, "arXiv preprint arXiv:1401.0640", (**2014**).

[11] 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, (**2010**).

[12] A. D. Diana Inkpen, "Canadian Acoustics / Acoustique canadienne", vol. 32, no. 3 (**2004**), pp. 130-131.

[13] F. Sebastiani, "ACM Comput. Surv.", vol. 34, no. 1, (**2002**). p. 1-47.

[14] W. M. Darling, "A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling", in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (**2011**).

[15] D. M. Blei, "Commun. ACM", vol. 55, no. 4, (**2012**), pp. 77-84.

[16] M. Steyvers and T. Griffiths, "Handbook of latent semantic analysis", vol. 427, no. 7,(**2007**), pp. 424-440.

[17] A. R. Mtafya, H. Dongjun, and G. Uwamahoro, "International Journal of Multimedia and Ubiquitous Engineering ", vol. 9, no. 12, (**2014**), pp. 97-106.

[18] M. Hoffman, D. M. Blei, and F. Bach, "Advances in Neural Information Processing Systems", vol. 23 (**2010**). pp. 856-864.

[19] [19].R. Řehůřek. 2014 [cited 2014 11-Dec]; Available from: http://radimrehurek.com/2014/03/tutorial-on-mallet-in-python/.