Keyword Extraction from Hindi Documents Using Statistical Approach

Aditi Sharan, Sifatullah Siddiqi and Jagendra Singh

Abstract Keywords of a document give us an idea about its important points without going through the whole text. In this paper, we propose an unsupervised, domain-independent, and corpus-independent approach for automatic keyword extraction. The approach is general and can be applied to any language. However, we have tested the approach on Hindi language. Our approach combines the information contained in frequency and spatial distribution of a word in order to extract keywords from a document. Our work is specially significant in the light that it has been implemented and tested on Hindi which is a resource poor and underrepresented language.

Keywords Keyword extraction · Spatial distribution · Standard deviation · Frequency · Hindi

1 Introduction

Our need to quickly sift through large amount of textual information is growing on a daily basis at an overwhelming rate. This task can be made easier if we have a subset of words (keywords) which can provide us with the main features, concept, theme, etc., of the document. Keyword extraction is also an important task in the field of text mining. Keywords can be extracted either manually or automatically but the former approach is very time-consuming and expensive. Automatic approaches for keyword extraction can be language-dependent or statistical

A. Sharan (✉) · S. Siddiqi · J. Singh
Jawaharlal Nehru University, New Delhi, India
e-mail: aditisharan@gmail.com

S. Siddiqi
e-mail: sifatullah.siddiqi@gmail.com

J. Singh
e-mail: jagendrasngh@gmail.com

L.C. Jain et al. (eds.), *Intelligent Computing, Communication and Devices*,
Advances in Intelligent Systems and Computing 309,
DOI 10.1007/978-81-322-2009-1_57

approaches, which are language independent. Most of the work in this area has been done in English language. Other languages are underrepresented due to scarcity of domain-specific resources and non-availability of data, especially labeled data. Now due to availability of large amount of textual data in different languages, there is a growing need for developing keyword extraction techniques in other languages as well.

There are many approaches by which keyword extraction can be carried out, such as supervised and unsupervised machine learning, statistical methods, and linguistic ones. Statistical methods for the extraction of keywords from documents have certain advantages over linguistic-based approaches such as the same approach can be applied to many different languages without the need to develop different set of rules each time for a different language. However, most of the statistical approaches are based on the corpus statistics. An obvious disadvantage of such methods is that they cannot be applied in the absence of availability of corpus. There are some other disadvantages as well of corpus-oriented keyword extraction methods. Keywords which occur in many documents of the corpus are less likely to be statistically discriminating, and most of the corpus-oriented approaches typically work on single words which are used in different contexts with different meanings. As availability of the corpus is a severe limitation in resource poor languages, it is important to focus on document-based approaches which can extract keywords from single document. Such methods are especially suitable for live corpuses such as news and technical abstracts. **Therefore, there is a need to find approaches which are unsupervised, domain independent, and corpus independent**.

Some work has been done for extracting keywords using statistical measures. Among one of the frequently used statistical measures to judge the importance of a particular word in the document is the frequency of a word in the document [1]. The intuition behind this measure is that the more important a word is to a document, the more number of times it should occur in the document. But in a document, most of the words having highest frequencies are stopwords (the, and, of, it, etc.). It was proposed in [2] to select the intermediate frequency words as keywords and discard the high- and low-frequency words as stopwords. Tf-idf measure [3] gives much better results than frequency score, but it requires a corpus and cannot be used on a single document. Shannon's entropy measure [4] was used to extract keywords from literary text in conjunction with randomly shuffled text. Spatial distribution [5] of words in the document was utilized to estimate the importance of a word in the document.

In this paper, we present a statistical approach for keyword extraction from single documents based on frequency and next nearest neighbor analysis of words in the document. We present an algorithm for the same. The algorithm has been tested on a famous Hindi novel "Godan" by Munshi Premchand.

2 Framework

This paper is based on the observation that both the frequency and spatial distribution of a word play an important role in evaluating its importance. Therefore, we have tried to develop an integrated approach which combines the information contained in frequency and spatial distribution of a word in order to extract keywords from a document. Our approach has strong theoretical background as well as empirical evidence. It works better than simple frequency-based approach as well as better than the approach based on spatial distribution only.

It has been observed that in a long text, the words with middle range of frequency are more important, because very high-frequency words are stopwords, whereas very low-frequency words are not important. We start with removing very low-frequency words.

Our spatial distribution-based approach is based on the observation that occurrence pattern for important words (keywords) should be different from that of non-important words. For a non-important word, its word distribution pattern should be random in the text and significant clustering should not be observed, whereas for a keyword, its distribution pattern should indicate some level of clustering because a keyword is expected to be repeated more often in specific contexts or portions of text.

To estimate the degree of clustering or randomness in the text for different words, we can use the standard deviation which measures the amount of dispersion in a data series. One such data series which can be analyzed is the successive differences between positions of occurrence of the word in the text. In other words, if a word **W** occurs **N** times in the document at positions X_1, X_2, X_3,…, X_N, then successive differences are $(X_2–X_1)$, $(X_3–X_2)$,…. $(X_N–X_{N-1})$. Representing the intermediate difference series for word W as $\mathbf{S_W}$ we have,

$$\mathbf{S_W} = \{(\mathbf{X_2} - \mathbf{X_1}),\ (\mathbf{X_3} - \mathbf{X_2}),\ (\mathbf{X_4} - \mathbf{X_3}), \ldots (\mathbf{X_N} - \mathbf{X_{N-1}})\}$$

For a word W with frequency N in the text, we have $N-1$ elements in the series S_W.

To eliminate the effect of frequency on the standard deviation analysis of different words, it is convenient to normalize the standard deviation ($\boldsymbol{\sigma_W}$) of series S_W with its corresponding mean ($\boldsymbol{\mu_W}$), so that normalized standard deviation of series S_W is $\hat{\sigma}_w$. Higher values of generally represent more pronounced clustering or lesser random behavior which is what we expect for important words.

Figures 1 and 2 show two words from the document (Godan) with similar frequencies, while the first word is a keyword, the other is a stopword. The clustering is evident in Fig. 1, while in Fig. 2, it can be easily seen that distribution is random throughout and no significant clustering is observed. As observed, spatial distribution can be used for differentiating between a keyword and a non-keyword.

Fig. 1 Spatial distribution of the keyword "होरी" with frequency 623 in the document

Fig. 2 Spatial distribution of stopword "गया" with frequency 677 in the document

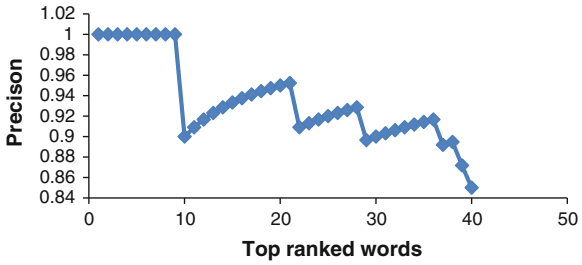

Fig. 3 Precision curve for *top* 40 words ranked on *decreasing* standard deviation. X-axis represents the ranks of words. The *curve* shows the number of keywords found in *top N* words ranked on $\hat{\sigma}$

2.1 Algorithm

1. Generate the list of unique words in the document which forms the vocabulary.
2. Calculate the frequency (f) of each unique word in the document.
3. Eliminate low-frequency words.
4. Generate the dataset of next nearest neighbor distances series $\mathbf{S_W}$ for each unique word in the document.
5. Calculate the mean ($\boldsymbol{\mu_W}$) and standard deviation ($\boldsymbol{\sigma_W}$) of dataset $\mathbf{S_W}$ for each unique word.
6. Normalize the $\boldsymbol{\sigma_W}$ with $\boldsymbol{\mu_W}$.
7. Rank the resulting words list in order of decreasing normalized standard deviation ($\hat{\sigma}_w$).
8. Select the words with highest standard deviation as keywords.

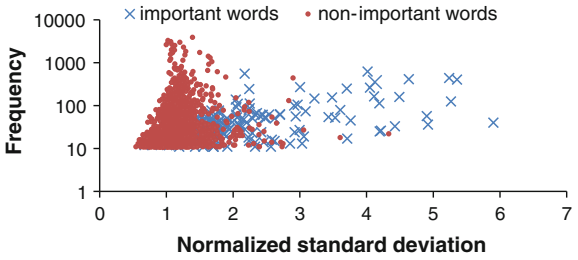

Fig. 4 Plot of frequency versus standard deviation of words in the document. *Frequency* has been plotted on logarithmic scale, whereas *normalized standard deviation* is plotted on linear scale. It can be seen that for $\hat{\sigma}$ >3, most of the words are important words

Table 1 *Top* ranked words from the document and their standard deviation scores

Keyword	Frequency (*f*)	$\hat{\sigma}$
चौधरी	40	5.899899
मेहता	404	5.35546
सलिया	125	5.26475
मालती	440	5.237255
नोहरी	57	4.902299
गोबर	414	4.633352
मङ्जिा	159	4.490566
हीरा	113	4.190991
धनिया	355	4.125608
गाय	164	4.112118
खन्ना	262	4.050105
होरी	623	4.015021
संपादक	45	3.760877
झुनिया	249	3.703393
गोवर्दि	79	3.604878
ओंकारनाथ	52	3.528352
भोला	155	3.48224
तंखा	54	3.42949
सोना	146	3.217998
मातादीन	74	3.085875

Table 2 Some stopwords from the document and their standard deviation scores

Stopword	Frequency (f)	$\hat{\sigma}$
हैं	1417	1.627264
था	1735	1.498282
है	3914	1.390569
ने	1841	1.289441
वह	1519	1.276871
नहीं	2306	1.255755
तो	3001	1.201025
की	2532	1.124765
न	1971	1.120748
भी	1711	1.091106
हो	2021	1.090015
को	2012	1.082808
के	2684	1.080228
का	1994	1.074362
और	3113	1.064317
ही	1241	1.063947
पर	1577	1.060343
कर	3053	1.027002
मे	3311	1.022517
से	2640	0.997308

3 Experiment

We performed our experiment on a novel "Godan" of Hindi language by Premchand. Some document statistics are as follows:

Total number of words in document = 167,707.

Total number of unique words in the document = 11,160.

Number of words with frequency greater than 10 = 1,565.

Words with frequency lesser than or equal to 10 were removed from consideration.

The proposed algorithm was implemented. The result is shown through a precision curve for top ranked 40 words in Fig. 3. It can be easily visualized that top ranked 10 words are all keywords of "Godan."

The result can be better understood with the help of Tables 1 and 2 and Fig. 4. Table 1 gives the list of top ranked words on $\hat{\sigma}$. Though many words have large frequency differences between them, but they have much closer values of standard deviation $\hat{\sigma}$. Table 2 gives a list of stopwords used in the document. Although these

words have quite larger frequencies than the important words in the document, their $\hat{\sigma}$ values are lower compared to the latter. Thus, $\hat{\sigma}$ works better in differentiating between important and non-important words. Figure 4 shows the distribution of words of "Godan" on frequency and $\hat{\sigma}$ where important and non-important words with frequency greater than 10 are labeled separately.

4 Conclusion

In this paper, we have proposed a novel document-based statistical approach to extract important words from Hindi literary document. Our approach hybridizes the information from both the frequency value and spatial distribution of words in the document. Standard deviation of nearest next neighbor distances of the word in the document was used as a discriminating factor. It was found that higher values of standard deviation generally correspond to the important words in the document. Considering that our approach is unsupervised, domain independent, and corpus independent, the results are quite motivating. Further, it was also observed that most of the important words which were extracted were named entities (NEs). Thus, a further research area which can be explored is the application of statistical methods to extract important named entities from literary texts.

References

1. Salton, G., Buckley, C.: Weighting approaches in automatic text retrieval. Inf. Process. Manage. **24**(5), 513–523 (1988)
2. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM J. Res. Dev. **1**(4), 309–317 (1957)
3. Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Documentation **28**(1), 11–21 (1972)
4. Herrera, J.P., Pury, P.A.: Statistical keyword detection in literary corpora. Eur. Phys. J. B. **63**(1), 135–146 (2008)
5. Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., Somoza, A.M.: Keyword detection in natural languages and DNA. Europhys. Lett. **57**, 759–764 (2002)