Research Article

International Journal of Advanced Computer Research, Vol 6(25)
ISSN (Print): 2249-7277   ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2016.625003

Keyword extraction from single documents using mean word intermediate distance

Sifatullah Siddiqi* **and Aditi Sharan**
School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

Received: 25-May-2016; Revised: 10-July-2016; Accepted: 19-July-2016

Abstract

Keyword extraction is an important task in text mining. In this paper a novel, unsupervised, domain independent and language independent approach for automatic keyword extraction from single documents have been proposed. We have used the word intermediate distance vector and its mean value to extract keywords. We have compared our approach with results from the standard deviation of intermediate distances approach as standard and found that there is heavy overlapping between the results of both approaches with the advantage that our approach is faster, especially in case of long documents as it removes the need to compute the standard deviation of word intermediate distance vector. Two famous works viz. "Origin of Species" and "A Brief History of Time" to demonstrate the experimental results have been used. Experiments show that the proposed approach works almost as better as the standard deviation approach and the percentage overlap between top 30 extracted keywords is more than 50%.

Keywords

Keyword extraction, Word means intermediate distance, Clustering, Standard deviation.

1.Introduction

Now day's large amounts of textual information are available to us through the internet and it is growing day by day. There is a pressing need to analyze these rapidly growing heaps of documents for various purposes. Keywords provide us with an efficient way to achieve this goal as they indicate main features, concept, theme etc. of a document. Keywords can be assigned either manually or automatically, but the former approach is very time-consuming and expensive and the need for automated processes that extracts keywords from documents is self-explanatory.

Keyword extraction is an important task in the field of text mining. There are many approaches by which keyword extraction can be carried out with each having its own pros and cons, but broadly speaking, there are four major methods as pointed out in [1] [2]:

1. Rule based linguistic approaches: These approaches are generally rule based and are derived from the linguistic knowledge/features. These can be more accurate, but are computationally intensive and require domain knowledge in addition to language expertise.
2. Statistical approaches: These approaches are generally based on linguistic corpus and statistical features derived from the corpus. The most important advantage of them is that they are independent of the language on which they are applied and hence the same technique can be used in multiple languages. These methods may not give as accurate results compared to linguistic ones, but the availability of large amounts of datasets has made it possible to perform statistical analysis and achieve good results.
3. Machine learning approaches: Machine Learning approaches generally employ supervised learning methods. In these methods keyword is extracted from training documents to learn a model and the learned model is tested through a testing dataset. After a satisfactory model is built it was used to find keywords from new documents. This approach includes Naïve Bayes, Support Vector Machine, etc. However, supervised learning methods for keyword extraction require a tagged document corpus, which is difficult to build. In the absence of such a corpus, we employ unsupervised and semi-supervised learning methods.
4. Domain specific approaches: Various approaches can be applied to domain specific corpuses

*Author for correspondence

which exploit the backend knowledge related to the domain (such as ontologies) and inherent structure of that particular corpus to identify and extract keywords.

2.Literature review

Major works in the field of keyword extraction using different approaches as outlined in the previous section have been reported by Siddiqi et al. [2]. Some of the other related works are as follows:

Sparck [3] proposed inverse document frequency, which is ubiquitous in term weighting schemes. Term frequency–inverse document frequency (TF-IDF), which involves multiplying the IDF weight by a TF weight, has proved very robust and difficult to beat. Salton et al [4] discussed various term weighting measures which are used most often along with their normalization factors. Buckley [5] stressed on the fact that proper weighting methods are very important and good weighting methods are more essential than feature selection process and both need to be handled simultaneously to be effective.

Researchers have used supervised learning techniques for extracting keywords. Turney [6] treated the issue of automatically extracting keyphrases as supervised learning task. It treats the document as a set of phrases which the learning algorithm classifies as positive or negative examples. Learning is performed with the help of C4.5 decision tree induction algorithm. Frank et al. [7] recommend the use of machine learning techniques and argued that it greatly improves the quality of automatic keyword extraction. Also, it creates domain-specific models from sets of training documents, which suitably modifies the judgments the model makes according to the set of documents from which is it extracting. Hulth [8] suggested that linguistic properties of texts yield higher quality keywords and better retrieval, and examines some different methods to include linguistic information into keyword extraction. Three methods of extraction are evaluated: n-grams, noun phrase (NP) chunks, and part-of-speech pattern matches. Terms are rated as keywords based on three features: document frequency, collection frequency, and relative position of its first occurrence in a document. Zhang et al. [9] employed conditional random field (CRF) model to extract keywords. CRF model is a new probabilistic model for segmenting and labelling sequence data. CRF is an undirected graphical model that encodes a conditional probability distribution with a given set of features. Litvake et al. [10] proposed DegExt, which is an unsupervised, graph-based, cross-lingual keyphrase extractor. DegExt uses a graph representation based on the simple graph-based syntactic representation of text and web documents, which enhances the traditional vector-space model by taking into account some structural document features.

Some researchers have also suggested the use of mathematical models and distributions for identifying stopword and keyword. Harter [11] examined the efficiency of a mixture of two Poisson distributions put forward by Bookstein et al. [12] to model the distribution of specialty words in the document collection. He argues that distribution of non-specialty words (non-keywords) lacks any order or structure in the text and is thus random. Ortuño et al. [13] demonstrated that important words of a text have a tendency to attract each other and form clusters. He argues that the standard deviation of the distance between successive occurrences of a word is such a parameter to quantify this self-attraction. Herrera et al. [14] tackled the problem of finding and ranking the relevant words of a document by using statistical information referring to the spatial use of the words. Shannon's entropy of information was used for automatic keyword extraction. The randomly shuffled text was used as a standard and the various measures used in the original document text were normalized by corresponding measures of random text.

Feng et al. [15] proposed an algorithm based on sequential patterns applied to a document which is represented as sequences of words. Important sequential patterns are extracted which reflect the semantic relatedness between words. Statistical as well as pattern features within words were used to build the keyword extraction model. The algorithm is language independent and does not require a semantic dictionary to get the semantic features. Hong et al. [16] proposed an improved keyword extraction method (extended TF). They used linguistic features of keywords like word frequency, part of speech, syntactical function of words, location appeared & word's morphology. On the base of the characteristics of each feature, weights were ascribed to different features and the support vector machine (SVM) model was used for further optimization.

Mehri et al. [17] described a method for ranking, the words in texts by use of non-extensive statistical mechanics. The non-extensively measure can be used to classify the correlation range between word-type occurrences in a text. C. Carretero-Campos et al. [18] improved upon the entropic and clustering approaches and proposed new metrics to evaluate the performance of keyword detectors to use them to find out the best approach of the two. It was observed that in general word clustering measures perform at least as well as the entropic measure, which requires a suitable partitioning of the text and word-clustering measures are also better for short texts since these measures discriminate better the degree of relevance of low frequency words than the entropic approach.

3.Proposed approach for keyword extraction

We proposed a novel and computationally efficient approach for keyword extraction from a single document by a simple estimation of the extent of clustering occurring in the usage of a particular word throughout the document. The motivation behind our approach was to build an unsupervised, language independent and domain independent method for keyword extraction.

It is a well-known fact, now that important words in a text are not randomly distributed, but they are instead clustered in certain regions of text where they appear with increased frequency and at relatively short distances while the words of no significance, such as stopwords, and non-keywords are almost randomly distributed throughout the text and don't exhibit pronounced clustering in their occurrence pattern. Thus, more the clustering is observed for a word, the more likely it is an important word for that document and vice versa. One approach to estimate this clustering exhibited by a word in a text was proposed in a seminal work [11] in which words with higher standard deviation of intermediate distances between the occurrences of a word, were generally seen to correspond to the keywords of the document.

Our approach eliminates altogether the need to calculate the standard deviation of intermediate distances and thereby improves the computational efficiency of the approach in [11] considerably.

3.1Theory of our approach

It's a known mathematical fact that the mean of a series of values lies between the largest and smallest values present in the series. So in a data series, there is a clustering of values about those points which are lesser than the mean and more dispersion of values about points which are larger than the mean. Now we assert the proposition that for a word the larger the number of values in the intermediate distance series, which are lesser than the mean value of the series, the more is the extent of clustering in that series and hence more important that word is in that document. Thus, by computing the fraction (f) of number of values in a data series, which are lesser than mean we can estimate the amount of clustering present in it. So, a way to estimate the relative order of clustering present in a multiple set of data series is to rank them in order of decreasing fractional values (f).

To estimate the importance of a word in the document we can analyze its set of successive differences between the positions of its occurrence in the text. In other words, if a word 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 S_W we have,

$S_W = \{(X_2–X_1), (X_3–X_2), (X_4–X_3), ... (X_N–X_{N–1})\}$

Mean of sequence S_W is,

$$\mu = \frac{(X_2-X_1)+(X_3-X_2)+\quad +(X_N-X_{N-1})}{N-1} = \frac{(X_N-X_1)}{N-1}$$

If the K number of elements in set S_W is lesser than μ, then the required fraction in f = K / (N-1).

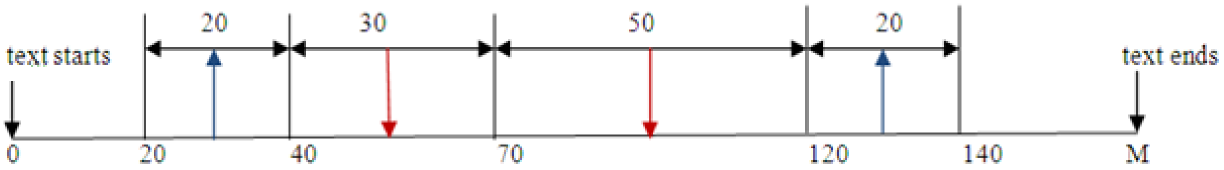

Figure 1 An example word occurrence pattern in a text

For example, let there be a total of M number of words in the document which are numbered successively along the text and a particular word say W occurs 5 times in the text at positions 20, 40, 70, 120 and 140 as shown in the *Figure 1*. Then the intermediate distance series is as follows:

S_W = {(40-20), (70-40), (120-70), (140-120)} = {20, 30, 50, 20}

Mean of S_W = (20+30+50+20) /4=120/4 = 30

We mark those distances which are lesser than the mean with blue upward pointing arrows and distances greater than or equal to mean with red downward pointing arrows at the midpoint of those distances. Let the number of upward pointing arrows be U and those of downward pointing is D. Then our required fraction is f = U/ (U+D)**.** A high value of fraction f indicates the high importance of the word under consideration.

By calculating the mean of sequence S_W and fraction of values which are lesser than mean we can have an estimate of the importance of a word in the document. For a set of words each having its own data series of intermediate distances we can rank them in decreasing order of their fraction of values which are lesser than mean and thus can have an estimate of the relative importance of the words of the document. We have selected two famous works viz. "Origin of Species" by Charles Darwin and "A Brief History of Time" by Stephen Hawking to demonstrate the results of our analysis. We show the spatial distribution for a keyword and a non-relevant word with similar frequencies respectively, for the document "Origin of Species" (*Figures 2 and 3*). While *Figures 4 and 5* show the same for the document "A Brief History of Time". The numbers along the axis represent the position of a word along the text. The vertical dark black lines in the upper part of each figure show the position of the word in the text while in the lower part of the figure solid blue line with arrows pointing upward are drawn at the midpoint of those intervals which are smaller than µ and dashed red lines with arrows pointing downwards are drawn at the midpoint of those intervals which are larger than µ. The ratio of the number of blue lines with the total number of blue and red lines is our required fraction.

These figures indicate the extent of clustering happening for different kinds of words in the document.

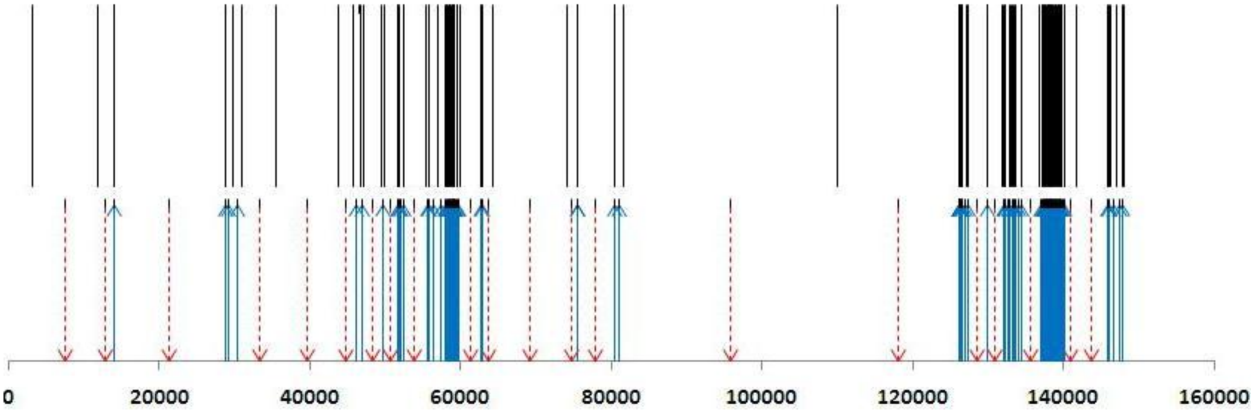

Figure 2 Spatial distribution of the keyword "**Organs**" with frequency 133 in "Origin of Species". Fraction of values lesser than the mean is 0.820

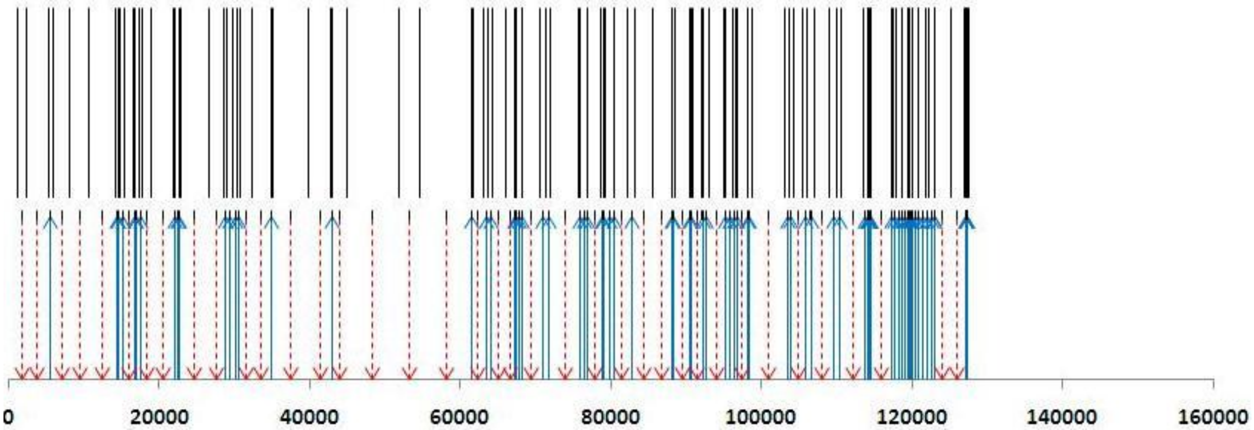

Figure 3 Spatial distribution of the non-relevant word "**Found**" with frequency 129 in "Origin of Species". Fraction of values lesser than the mean is 0.664

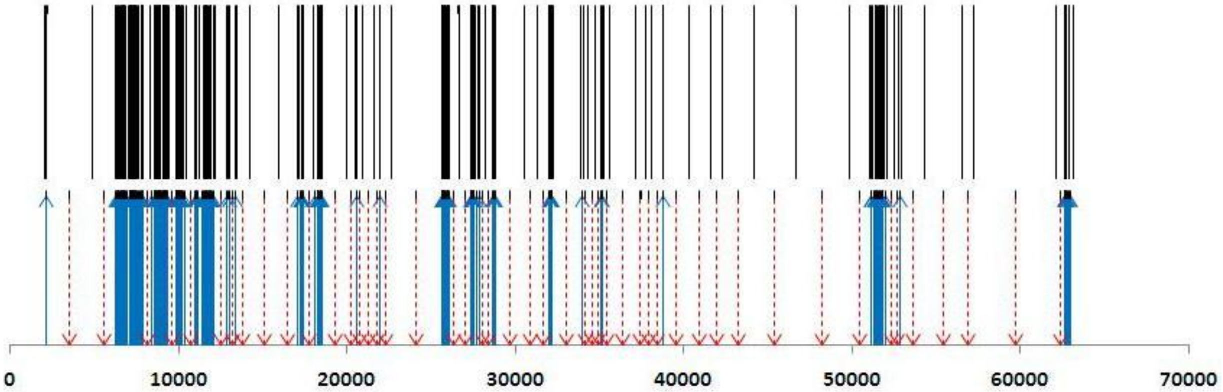

Figure 4 Spatial distribution of the keyword word "**Light**" with frequency 243 in "A Brief History of Time". Fraction of values lesser than the mean is 0.802

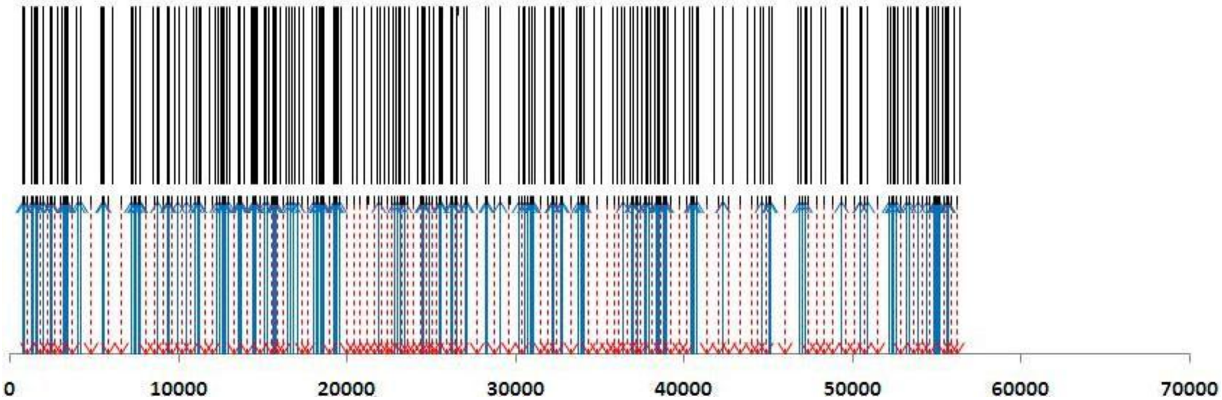

Figure 5 Spatial distribution of the non-relevant word "**Other**" with frequency 265 in "A Brief History of Time". Fraction of values lesser than the mean is 0.633

3.2Preparing the documents

We have selected two famous works viz. "Origin of Species" by Charles Darwin and "A Brief History of Time" by Stephen Hawking to demonstrate the results of our analysis. All the punctuation marks were removed from both documents as well as everything besides the main text of the documents such as table of contents, appendix, comment on edition of books, index and glossary were removed. The whole text was converted into lowercase and singular and plural forms of a word were treated as different words and no stemming was performed on the texts. As a result, we have 149111 total no of words and 7254 word forms in "Origin of Species" and a total of 63505 words and 4488 word forms in "A Brief History of Time".

4.Experimental results

For the document "Origin of Species" *Table 1* shows the top 30 words extracted by our approach along with their fractional values (f) while *Table 2* shows the top 30 words extracted via standard deviation approach [11].

Table 1 Top keywords extracted from "Origin of Species" via fraction of mean intermediate distance approach

Word	Fraction	Word	Fraction
Slaves	0.969697	Ants	0.916667
Bees	0.944444	Lowlands	0.916667
Deposits	0.944444	Barriers	0.916667
Floated	0.941177	Fertility	0.910256
Comb	0.941177	Flat	0.909091
Instincts	0.939394	Heath	0.909091
Gartner	0.933333	Instinct	0.909091
Pollen	0.930556	Analogical	0.909091
Sterility	0.930556	Instinctive	0.909091
Hybrids	0.930435	Warmer	0.909091
Wall	0.928571	Formations	0.909091
Workers	0.923077	Hive-bee	0.904762
Systems	0.923077	Transport	0.904762
Basins	0.916667	Temperate	0.903226
Construction	0.916667	Bars	0.9

Table 2 Top keywords extracted from "Origin of Species" via standard deviation of mean intermediated distance approach

Word	Std-Dev	Word	Std-Dev
Formations	5.476387	Gartner	3.695682
Bees	5.399093	Wall	3.512741
Hybrids	4.884996	Groups	3.445037
Sterility	4.819015	Continents	3.372515
Instincts	4.61384	Sterile	3.348341
Workers	4.478911	Glacial	3.342532
Slaves	4.464653	Breeds	3.317335
Diagram	4.214492	Basins	3.277944
Instinct	4.116471	Sexual	3.25111
Island	4.112764	Nest	3.243127
Ants	4.066406	Shores	3.22787
Fertility	4.051916	Fertile	3.223329
Pollen	4.024308	Value	3.199455
Organ	3.91079	Homologous	3.175496
Cells	3.877882	Striped	3.160831

The running percentage overlap between the top 30 words generated by the two algorithms for "Origin of Species" is shown in *Figure 6* while the same for "A Brief history of Time" is shown in *Figure 7*. It represents the number of common keywords found in both approaches at each successive rank.

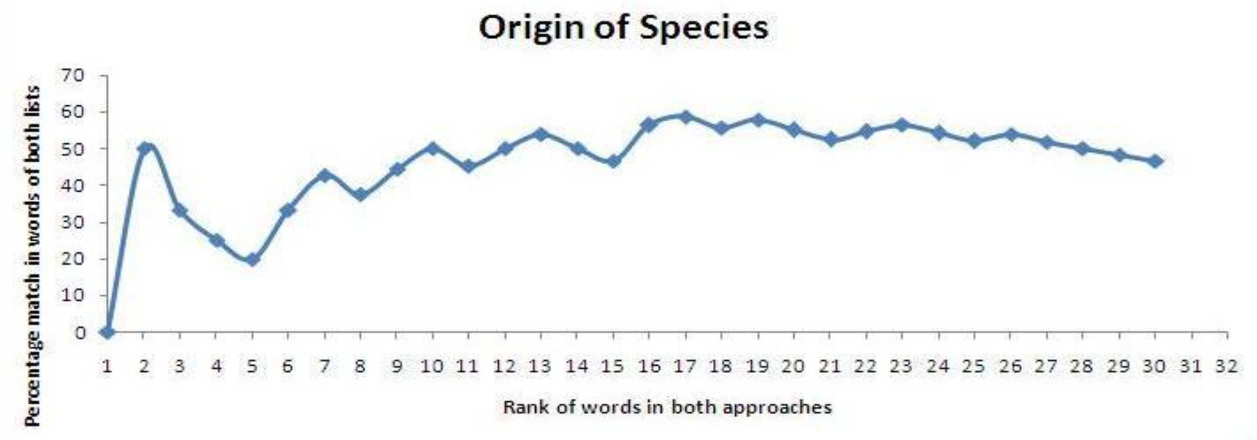

Figure 6 Running percentage match between top keywords extracted from two approaches for "Origin of Species"

For the document "A Brief History of Time" *Table 3* shows the top 30 words extracted by our approach along with their fractional values (***f***) while *Table 4* shows the top 30 words extracted via standard deviation approach [11].

Table 3 Top keywords extracted from "A Brief History of Time" via fraction of mean intermediate distance approach

Word	Fraction	Word	Fraction
Disorder	0.971429	Ask	0.9
Friedmann	0.967742	Bubbles	0.9
Arrow	0.941177	Centauri	0.9
Area	0.928571	Table	0.9
Coordinates	0.923077	Temperature	0.897436
Plates	0.923077	Virtual	0.896552
Quark	0.916667	Condition	0.888889
String	0.913044	Histories	0.885714
Decay	0.909091	Black	0.885106
Box	0.909091	Moon	0.882353
Gamma	0.904762	Spin	0.878788
Miles	0.904762	Imaginary	0.875
Quarks	0.902439	Scientific	0.875
Meter	0.9	Entropy	0.875
Necessary	0.9	Backward	0.875

Table 4 Top keywords extracted from "A Brief History of Time" via standard deviation of mean intermediated distance approach

Word	Std-Dev	Word	Std-Dev
Disorder	4.733637	Plates	3.34487
String	4.689268	Area	3.291883
Friedmann	4.629699	Event	3.196331
Arrow	4.057601	Histories	3.137497
Imaginary	3.910357	Box	3.035857
Entropy	3.835296	Quark	3.01141
Black	3.551782	Electrons	3.005598
Newton	3.526672	Twice	3.003511
Quarks	3.502943	Galileo	2.99651
Hole	3.472643	Exclusion	2.987039
Coordinates	3.433051	Chandrasekhar	2.953399
Dimensional	3.38156	Star	2.943367
Temperature	3.374085	Curved	2.931851
Dimensions	3.373313	Phase	2.907753
Primordial	3.349099	Bubbles	2.906903

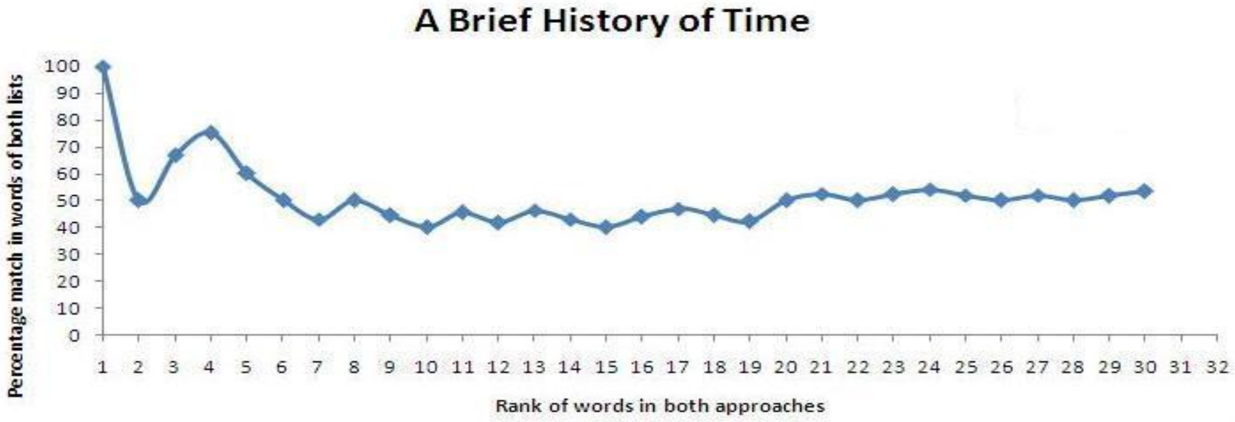

Figure 7 Running percentage match between top keywords extracted from two approaches for "A Brief History of Time"

The results show that there is a heavy overlapping between the results of the standard deviation approach and our mean distance value approach. The set of keywords returned by mean value approach is very similar to standard deviation approach for both the documents. The minor difference in results from standard deviation and mean distance value can be explained on the basis that the two approaches estimate the importance of a word along different dimensions. The advantage of our approach is that it does not require the computation of standard deviation of the intermediate distance vector and thus is simpler.

5.Conclusion

We have presented a novel approach for automatic extraction of keywords from single documents. In our approach we have used the set of intermediate distances of a word to calculate the mean intermediate distance for that word in the document and computed the fraction of those distances which are lesser than the mean intermediate distance. Words are ranked in decreasing order of fractional values. We have compared our results with the standard deviation approach and found that there is a considerable overlap between the results of both approaches. In future we would apply this algorithm to different domain texts and would be interested to observe the effectiveness of this algorithm on a different language text.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Zhang C, Wang H, Liu Y, Wu D, Liao Y, Wang B. Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems. 2008; 4(3):1169-80.

[2] Siddiqi S, Sharan A. Keyword and keyphrase extraction techniques: a literature review. International Journal of Computer Applications. 2015; 109(2):18-23.

[3] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. 1972; 28(1):11-21.

[4] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management. 1988; 24(5):513-23.

[5] Buckley C. The importance of proper weighting methods. In proceedings of the workshop on human language technology 1993 (pp. 349-52). Association for Computational Linguistics.

[6] Turney PD. Learning algorithms for keyphrase extraction. Information Retrieval. 2000; 2(4):303-36.

[7] Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-Manning CG. Domain-specific keyphrase extraction. In international joint conference on artificial intelligence 1999 (pp. 668-73).

[8] Hulth A. Improved automatic keyword extraction given more linguistic knowledge. In proceedings of the conference on empirical methods in natural language processing 2003 (pp. 216-23). Association for Computational Linguistics.

[9] Zhang C. Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems. 2008; 4(3):1169-80.

[10] Litvak M, Last M, Aizenman H, Gobits I, Kandel A. DegExt-A language-independent graph-based keyphrase extractor. In advances in intelligent web mastering–3 2011 (pp. 121-30). Springer Berlin Heidelberg.

[11] Harter SP. A probabilistic approach to automatic keyword indexing. Part II. An algorithm for

probabilistic indexing. Journal of the American Society for Information Science. 1975; 26(5):280-9.

[12] Bookstein A, Swanson DR. Probabilistic models for automatic indexing. Journal of the American Society for Information Science. 1974; 25(5):312-6.

[13] Ortuño M, Carpena P, Bernaola-Galván P, Muñoz E, Somoza AM. Keyword detection in natural languages and DNA. EPL (Europhysics Letters). 2002; 57(5):759-64.

[14] Herrera JP, Pury PA. Statistical keyword detection in literary corpora. The European Physical Journal B. 2008; 63(1):135-46.

[15] Feng J, Xie F, Hu X, Li P, Cao J, Wu X. Keyword extraction based on sequential pattern mining. In proceedings of the third international conference on internet multimedia computing and service 2011 (pp. 34-8). ACM.

[16] Hong B, Zhen D. An extended keyword extraction method. International conference on applied physics and industrial engineering 2012 (pp. 1120-7). Physics Procedia.

[17] Mehri A, Darooneh AH. Keyword extraction by nonextensivity measure. Physical Review E. 2011; 83(5):056106.

[18] Carretero-Campos C, Bernaola-Galván P, Coronado AV, Carpena P. Improving statistical keyword detection in short texts: entropic and clustering approaches. Physica A: Statistical Mechanics and its Applications. 2013; 392(6):1481-92.

Sifatullah Siddiqi is a research scholar at School of Computer and Systems Sciences at Jawaharlal Nehru University (JNU), New Delhi. His current research interests are in unsupervised and statistical keyword / keyphrase extraction techniques for documents. He did his M. Tech. in computer science from JNU and completed his B.Tech. in Computer Engineering from Zakir Hussain College of Engineering & Technology, Aligarh Muslim University (AMU), Aligarh.
Email: sifatullah.siddiqi@gmail.com

Aditi Sharan is an Assistant Professor at the School of Computer and Systems Sciences, Jawaharlal Nehru University. Her research interests include Text mining, Information retrieval and natural language processing.