World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:6, No:12, 2012

Examining the Value of Attribute Scores for Author-Supplied Keyphrases in Automatic Keyphrase Extraction

Vicky Min-How Lim, Siew Fan Wong, and Tong Ming Lim

***Abstract*—**Automatic keyphrase extraction is useful in efficiently locating specific documents in online databases. While several techniques have been introduced over the years, improvement on accuracy rate is minimal. This research examines attribute scores for author-supplied keyphrases to better understand how the scores affect the accuracy rate of automatic keyphrase extraction. Five attributes are chosen for examination: Term Frequency, First Occurrence, Last Occurrence, Phrase Position in Sentences, and Term Cohesion Degree. The results show that First Occurrence is the most reliable attribute. Term Frequency, Last Occurrence and Term Cohesion Degree display a wide range of variation but are still usable with suggested tweaks. Only Phrase Position in Sentences shows a totally unpredictable pattern. The results imply that the commonly used ranking approach which directly extracts top ranked potential phrases from candidate keyphrase list as the keyphrases may not be reliable.

***Keywords*—**Accuracy, Attribute Score, Author-supplied keyphrases, Automatic keyphrase extraction.

I. Introduction

AUTOMATIC keyphrase extraction is an automated process to annotate keyphrases for a particular document based on the document content itself [1–7]. It tries to mimic keyphrase annotation task of human experts [1–7]. With gazillion of documents currently stored in online databases that are without readily assigned keyphrases, it is an onerous if not impossible task to manually assign keyphrases to each and every document. Thus, an automated process is needed to assist in keyphrase annotation task.

Previous research has been conducted to automatically annotate keyphrases for documents [1–7]. However, the accuracy rate is still considered low. Accuracy for automatic keyphrase extraction is often measured using precision and recall rate. Precision rate is the measurement of the probability or ratio at which a generated keyphrase list matches the author-supplied keyphrase list [8]. Recall rate is the measurement of the probability or ratio at which an author-supplied keyphrase will be selected as the automatically generated keyphrase[1] [8].

These research reported low precision rate of around 0.30 out of 1.00 [2], [9] and low recall rate of around 0.50 out of 1.00 [1]. Ref. [1], [2] found that candidate list which contains the candidate keyphrases covers 60.62% and 72% of author-supplied keyphrases respectively. The final extracted keyphrase list contains only up to 50% of author-supplied keyphrases. This means that there is still about 10-20% of keyphrases that appear in the candidate list but fail to get to the top of the ranked candidate list. Therefore, improvement needs to be made to the weighting and ranking process of automatic keyphrase extraction in order to improve the overall accuracy rate.

The weighting and ranking of candidate phrases for automatic keyphrase extraction is solely based on the score of one or more attributes. These scores will affect the position at which a keyphrase is being ranked in the candidate list. A keyphrase that ranks at the top will be chosen during automatic keyphrase extraction process while a keyphrase that ranks at the bottom will not be considered. To the knowledge of the authors, previous research (e.g., [1–10]) only focus on comparing the generated keyphrase list to the author-supplied keyphrase list to obtain the final results of precision and recall rate. The goal is to just compute the overall accuracy rate. These are done without delving into the question of how different attributes relate to author-supplied keyphrases.

The objective of this paper is therefore to examine the relationship between different attribute scores and author-supplied keyphrases. The goal is to better understand how these attribute scores affect the precision and recall rate of automatic keyphrase extraction. Through these examinations, a better weighting formula can be developed to further improve the accuracy of automatic keyphrase extraction.

This paper is structured as follows. Section II reviews related work of automatic keyphrase extraction. Section III describes the methodology of the study while Section IV discusses the experimental results. Section V concludes the paper with suggestions for future research work.

II. Related Work

Different attributes have been used in the literature to identify keyphrases in automatic keyphrase extraction. The following briefly describes some commonly used attributes.

Vicky Min-How Lim is with the Faculty of Science & Technology, Sunway University, Bandar Sunway, 46150 Selangor Malaysia (phone: 0168768505; e-mail: minhow.lim@gmail.com).

Siew Fan Wong is with the Faculty of Science & Technology, Sunway University, Bandar Sunway, 46150 Selangor Malaysia (phone: +603-74918622 e-mail: siewfanw@sunway.edu.my).

Tong Ming Lim is with the Faculty of Science & Technology, Sunway University, Bandar Sunway, 46150 Selangor Malaysia (phone: +603-74918622 e-mail: tongmingl@sunway.edu.my).

[1] Both precision and recall rates are calculated based on an author-supplied keyphrase list with the assumption that the list is 100% accurate and complete.

Term Frequency (TF) calculates the number of times in which a particular phrase appears in a document [2], [3]. It is often used along with Inverse Document Frequency (IDF) as TF x IDF [1], [5], [9]. Given a candidate keyphrase, TF is the frequency at which a keyphrase occurs in the target document while IDF is the number of documents in the collection corpus that contains a given keyphrase [1]. IDF can be calculated based on either a domain corpus or a general corpus. When it is calculated based on a domain corpus, the IDF score indicates the rarity of a candidate keyphrase within the particular domain [1]. This works based on the assumption that rare information can help to boost the performance of domain specific automatic keyphrase extraction [4]. When it is calculated based on a general corpus such as Google n-gram counts, the IDF score indicates the rarity of a candidate keyphrase in general use. When IDF is calculated in such a way, it results in insignificant impact on the precision and recall rate and thus can be ignored [10]. When putting together TF and IDF, it measures the level of rarity of a target phrase in a domain or in general use [5]. There are a few ways to put them together: multiplication, summation, and division. Note that IDF can only be used if a corpus is provided for training or a word count dictionary such as Google n-gram count is available [9].

First Occurrence attribute indicates the position where a particular phrase first appears within a document [5], [6], [9], [11]. The assumption is that important phrases often appear early in a document. First Occurrence of a phrase is calculated as the number of terms that precedes a target phrase while Last Occurrence of a phrase is calculated as the number of terms that succeeds a target phrase [10]. First and Last Occurrence can also be represented in a percentage format to indicate the position of a phrase in a document. In GenEx [6], First Occurrence is used as a guideline to give reward or penalty score for a phrase. In KEA [4-5], it is used as one of the classifying factors along with TFxIDF to classify a phrase into either a keyphrase or a non-keyphrase. KP-Miner uses First Occurrence to filter phrases that are unlikely to be keyphrases.

Context word is a word that appears together with the targeted phrase in a sentence [3]. The underlying idea is that keyphrases tend to share more context words compared to other phrases.

Phrase position in Sentences identifies object phrases and subject phrases in a sentence based on grammatical facts that subject phrases come early in a sentence and object phrases come near the end of a sentence [2]. This information is used to further verify whether an n-gram is a noun phrase.

Phrase Length is the number of words in a phrase [1], [5], [6]. Based on common human bias toward a particular length of phrases when they choose keyphrases for a document, higher priority can be given to specific phrase length. However, since human preference is subjective and might vary across different countries, education backgrounds, fields of study, and occupation, there might be the need to re-specify the length of phrases for each domain corpus.

Term Cohesion Degree calculates how likely an n-gram is to be a phrase [12]. This is useful as an alternative attribute in the absence of noun phrase identifier. However, it is rarely used in previous work.

$$DegreeofTermCohesion = \frac{|T| \times \log_{10} f(T) \times f(T)}{\sum_{w_i \in T} f(w_i)}$$

where, T is the number of words in term T
f(T) is the frequency of term T
$f(w_i)$ is the frequency of word w_i in T

Boost factor is a generalized term for any factor that boosts up the ranking score of one candidate keyphrase. This is used in previous work to solve bias issue such as TF overlapping issue [6], [9] and IDF bias issue encountered in [9]. However, it faces similar issue as that in Phrase Length whereby the factor value might need to be re-specified for each domain corpus.

Table I shows a brief summary of the attributes used in the literature.

TABLE I
ATTRIBUTES USED IN WEIGHT CALCULATION

Research	Attributes						
	First Occurrence	Phrase length	TF only	TF x IDF	Phrase Position in sentences	Context word	Boost factor
GenEx by Turney [6]	√	√					√
KEA by Frank et al. [4][5]	√			√			
KP-Miner by El-Beltagy et al. [9]	√			√			√
C/NC-Value by Frantzi et al. [3]			√			√	
N-gram filtration by Kumar et al.[2]			√		√		
MLP by Sakar et al. [1]		√		√			

III. METHODOLOGY

After reviewing existing research work (e.g., [1–5], [8–10], [12], [13]), five attributes are chosen for testing in this study: TF, First Occurrence, Last Occurrence, Phrase Position in Sentences, and Term Cohesion Degree. The data set used is the Journal 2 dataset mentioned in [9]. In [9], KP-Miner was benchmarked with Extractor [6], [8], [13] and KEA [4], [5]. This means by using this dataset, the final results generated for precision and recall rate from this study can be compared not only with that of KP-Miner, but also with that of Extractor and KEA. This will help to increase the confidence toward the results produced from this study.

There are 60 documents in Journal 2 dataset. Within these 60 documents, an average of 93.35% of author-supplied keyphrases can be found in the respective documents. The five

attribute scores for the keyphrases that can be found in the respective documents are calculated and plotted into a series of graphs. These graphs give a clear view on the performance of the attribute scores across the 60 documents. Additionally, boundaries of attribute scores are examined to understand where author-supplied keyphrases are located in the ranked candidate keyphrase list for the purpose of ranking approach in automatic keyphrase extraction.

IV. Results and Discussion

The following presents the results of the experiments. A total of five experiments were carried out to gauge an idea of how each of the attribute performs.

Fig. 1 shows the results of experiment for TF. The x-axis is the label for each of the 60 documents examined. The y-axis is the average occurrence of a particular keyphrase in each document. For example, authors supplied four keyphrases for Document #6: "FOAF", "metadata", "keyword extraction", and "word cooccurrence". A count will be conducted for how many times "FOAF", "metadata", "keyword extraction", and "word cooccurrence" appear in Document #6. Assume that "FOAF" appears 68 times, "metadata" appears 57 times, "keyword extraction" appears 26 times, and "word cooccurrence" appears 1 time, then the average TF for Document #6 will be 38. This average frequency of 38 is plotted in Fig. 1. The same goes from the remaining 59 documents in the dataset.

From Fig. 1, it is evident that with the exception of a few spikes, the majority of the documents have an average frequency of ten. This means that the frequency at which author-supplied keyphrases appears in most of the 60 documents examined in this study is ten. When averaging across all the 60 documents, the mean frequency is 16.49. Considering the fact that a typical scientific document usually contains thousands of words ("journal 2" dataset has an average of 6640.8333 words across the 60 documents), an average of ten or even 16.49 shows surprisingly low occurrence of author-supplied keyphrases that actually are present in the respective documents. This means that previous studies (e.g., [2], [3]) which directly incorporate TF into the weighting and ranking process may not be that reliable. As for the few spikes that show up to 80 in average frequency, one potential cause is the differences in document length. The longer a document is, the higher the spike will be.

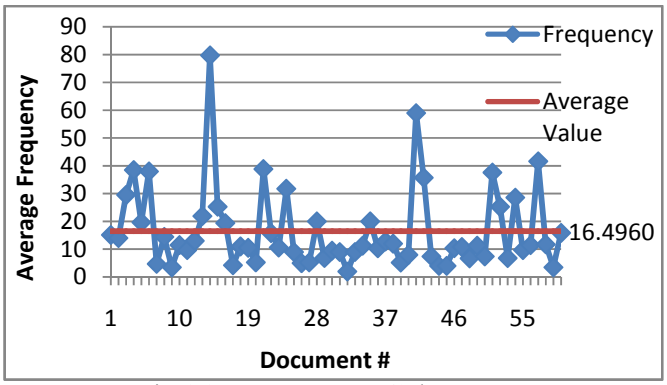

Fig. 1 Average TF across 60 documents

To remove the effect of document length, the ratio of TF over the document length is used. The ratio is calculated as TF divided by total word count in the document. Fig. 2 shows that while more variations appear across the dataset, these variations lie only within a small range (0.00 – 0.01). This means that author-supplied keyphrases mostly occupied less than 1% of the total words in a document. With this information, future processing of automatic keyphrase extraction can filter out those phrases that occupy more than 1% of the total words within a document.

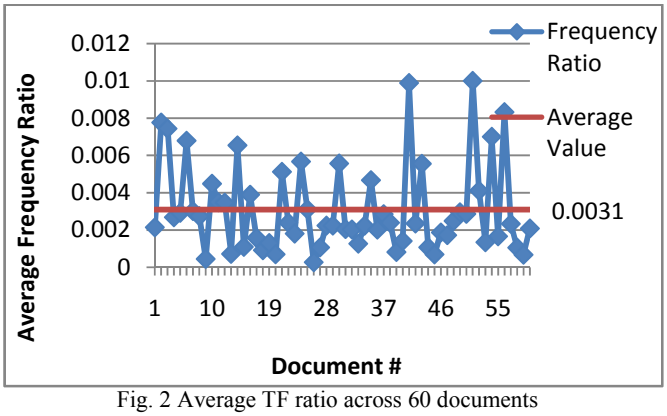

Fig. 2 Average TF ratio across 60 documents

Fig. 3 shows the average Term Cohesion Degree across the 60 documents. It is evident that Term Cohesion Degree does not have any consistent pattern and has a wide range of variation (0.01 to 1.51). This might be attributed to its dependency on TF in the formula calculation. The bias in TF that varies based on total length of a document affects the results of Term Cohesion Degree. As a result, the attribute becomes totally unpredictable and unreliable.

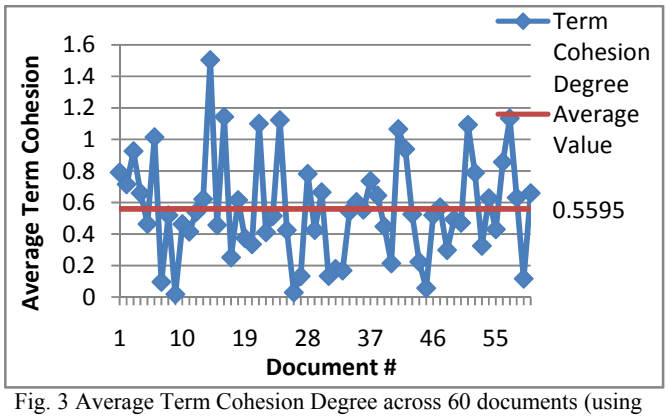

Fig. 3 Average Term Cohesion Degree across 60 documents (using TF)

Term Cohesion Degree is then re-examined by replacing TF with TF Ratio (see Fig. 4). The results are in negative values because part of the formula for Term Cohesion Degree "$\log_{10} f(T)$" will produce negative value whenever f(T) is a ratio. The range of variation has now increased to "0.0 – -2.6". This increase in variation reduces the reliability furthermore compared to the original formula that uses only TF.

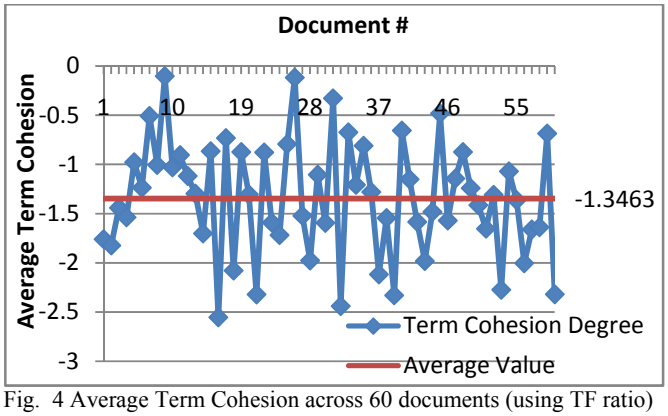

Fig. 4 Average Term Cohesion across 60 documents (using TF ratio)

First Occurrence, Last Occurrence, and Phrase Position in a Sentence are plotted using a scale of 0.00 to 1.00 whereby 0.00 indicates the first position at which an author-supplied keyphrase appears in a document or a sentence and 1.00 is the last position at which an author-supplied keyphrase appears in a document or a sentence. Fig. 5 shows the average First and Last Occurrence of author-supplied keyphrases across the 60 documents. The test of First Occurrence shows only a small range of variation (from 0.00 to 0.25). However, for Last Occurrence, the range is big (from 0.18 to 1.00). In some cases, for examples, Document #1 has a variation range of 0.44 - 1.00 (see Fig. 6), meaning that Last Occurrence can be present at any place within the second half of the document; Document #13 has a variation range of 0.01 - 1.00 (see Fig. 7), meaning that Last Occurrence can be present at any place throughout the entire document. However, these situations mostly occur when a particular keyphrase appears only once in a document. The data also shows that the average for First Occurrence is 0.05% which means author-supplied keyphrases usually appear around the first 5% of the document. Average Last Occurrence of 0.72% means most keyphrases also come in around the last 30% of the document.

In summary, First Occurrence is consistent and reliable to be used as an attribute for automatic keyphrase extraction. On the other hand, more investigations on Last Occurrence need to be conducted before it is used as an attribute in automatic keyphrase extraction. One suggestion is to disregard those phrases that occur only once in a document.

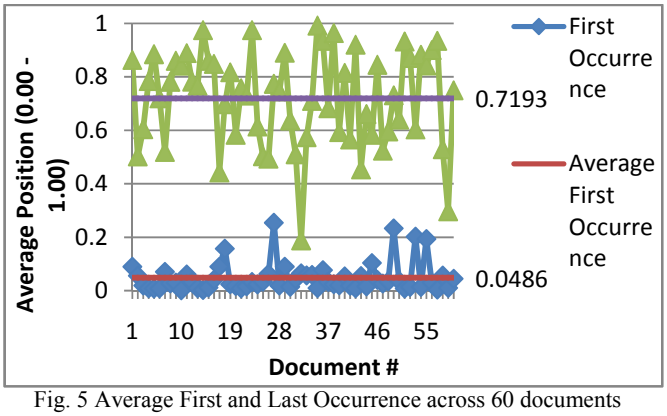

Fig. 5 Average First and Last Occurrence across 60 documents

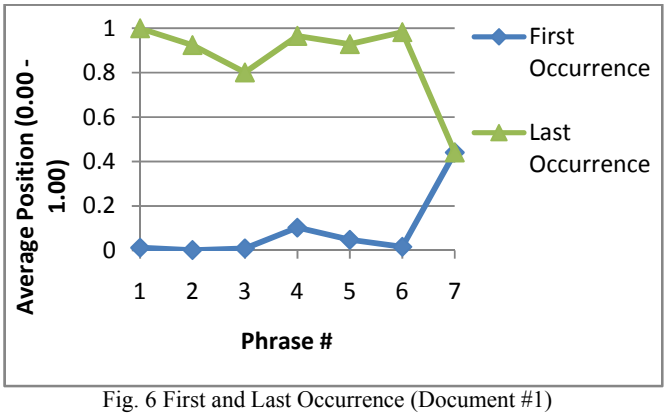

Fig. 6 First and Last Occurrence (Document #1)

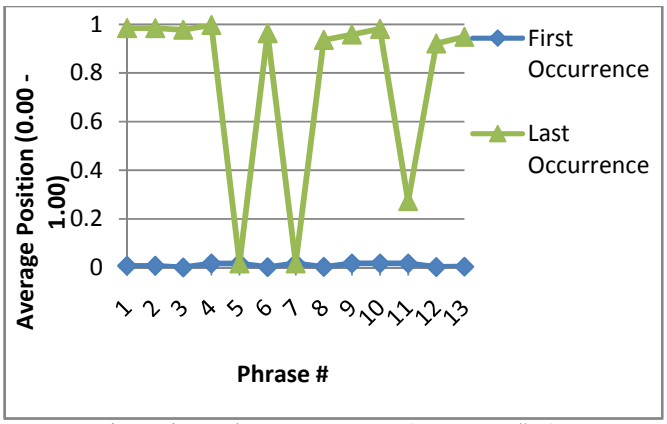

Fig. 7 First and Last Occurrence (Document #13)

Fig. 8 shows the results of Last Occurrence experiment that ignores keyphrases that appear only once in the document. Apparently, most of the average Last Occurrences concentrated at the end of the documents. Compared to Fig. 5, the range of variation reduces significantly. This confirms our

previous suggestion that Last Occurrence can serve as an attribute only after candidate keyphrases that occur only once in a document is ignored.

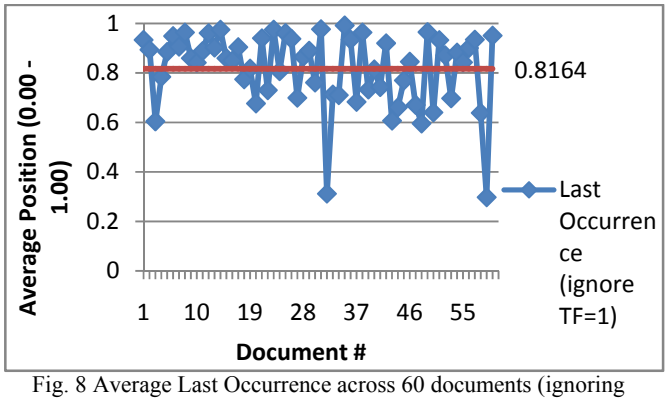

Fig. 8 Average Last Occurrence across 60 documents (ignoring TF=1)

Fig. 9 shows the average position of author-supplied keyphrases in sentences for each of the 60 documents. It gives information on whether a keyphrase tends to appear either toward the start of a sentence or the end of a sentence. As shown in the figure, there are some variations for minimum, maximum, average position of keyphrases in a sentence. Minimum position in a sentence identifies the average earliest position where a keyphrase appears in a sentence of a document while maximum position in a sentence is the average latest position where a keyphrase appears in a sentence of a document. Identification of the minimum or maximum position of a keyphrase in a sentence is based on grammatical rules where noun phrase usually comes earlier in a sentence [2]. Similarly, some object references that can be used as keyphrases may also appear toward the end of a sentence [2].

From the results, the average minimum value consistently occurs in the first half of a sentence while the average maximum value consistently occurs in the second half of a sentence. This shows that the attributes of average minimum and maximum position are useful in identifying keyphrases. However, the information on whether all keyphrases lay only at the beginning or at the ending part of a sentence remains unknown.

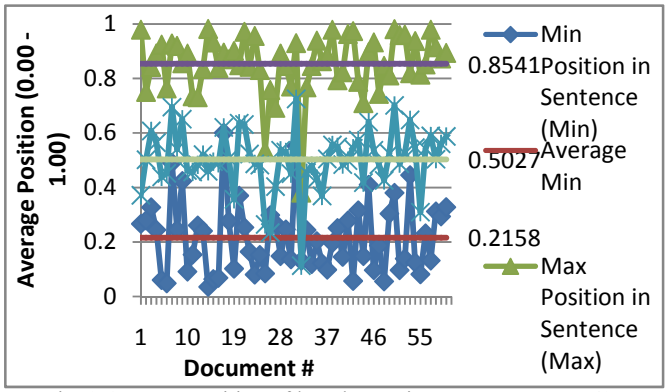

Fig. 9 Average position of keyphrases in sentences across 60 documents

In order to know whether all keyphrases occur either at the starting or the ending of a sentence, the detailed position information of keyphrases in sentences were examined. Fig. 10 gives an example of position information for three author-supplied keyphrases ("Consensys", "Paxos", and "two-phase commit") in Document #14[2]. From the figure, it is clear that there is no one consistent pattern for the positioning of author-supplied keyphrases. In fact, these keyphrases tend to randomly appear in any part of a sentence. This inconsistent positioning pattern contrasts English grammatical rules in identifying concept, subject, and object as argued in [2]. Therefore, the position of keyphrases in sentences cannot be used as a reliable attribute for automatic keyphrase extraction.

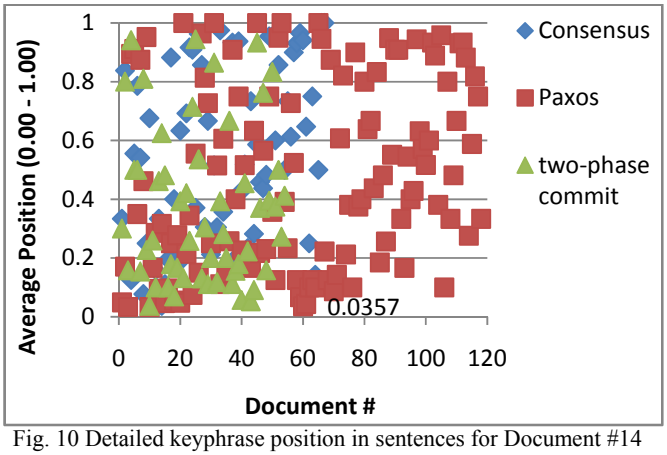

Fig. 10 Detailed keyphrase position in sentences for Document #14

Additionally, the boundaries of each attribute were calculated to obtain information on where author-supplied keyphrases might appear if all the words in a document are put into a ranked list. Table II summarizes the average boundaries for two attribute scores (TF and Term Cohesion Degree) examined in this paper. Other attribute scores (First Occurrence, Last Occurrence, and Position in Sentences) are not included because they have 0.00 for the lowest boundary and nearly 1.00 for the highest boundary.

In the first experiment, TF and Term Cohesion Degree of every word in each document were calculated to extract the highest and lowest boundaries of the attribute scores. In the second experiment, the first experiment was repeated except that now TF ratio was used instead of TF only. The third and fourth experiments repeated what had been done in the first and second experiments with an additional stopword[3] filter.

TABLE II
AVERAGE BOUNDARIES

Experiment	Average Frequency		Average Term Cohesion Degree	
	Highest	Lowest	Highest	Lowest
1	851.8167	1	2.8166	0
2	0.1262	0.000241	0	-3.7194
3	181.9500	1	2.1345	0
4	0.0288	0.000241	0	-3.7194

[2] Document #14 is chosen because it has the highest average frequency of keyphrases (see Fig. 1) and thus will give more data points for further examination.

[3] Stopwords include conjunctive words and pronoun words.

From the results, it is clear that direct ranking and extraction of top ranked candidate is unreliable because the average positions where these two attributes (TF and Term Cohesion Degree) fall are relatively far from the highest boundary. At the same time, they are also some distances away from the lowest boundary. Take for example, the average TF across 60 documents is 16.49 (Fig. 1). This number is far from the highest boundary of 851.8167 and 181.9500 (with stopwords filtered). Therefore, while TF displays a consistent small variation range, it is not practical to use only TF as a ranking score.

Similarly, the average TF ratio across 60 documents is 0.0031 (Fig. 2). This number is also relatively far from the highest boundary of 0.1262 and 0.0288 (with stopwords filtered). Furthermore, all author-supplied keyphrases only occupy less than 1% of the total word count in the corresponding documents. So, it is also not practical to use TF ratio directly as a ranking score to extract keyphrases from top ranked candidate list. Even if the list is ranked from the lowest to the highest, the value is still a distance away from the lowest boundary of 0.000241.

Nevertheless, since there is a consistently smaller range of variation for TF, the average score of the variation can be used as a reference for the most likely point to be keyphrases. TF score of each candidate keyphrase can then be converted into a likelihood rating whereby a candidate keyphrase score that is nearest to the reference point will be given a higher likelihood rate compares to a candidate keyphrase score that is further from the reference point. The same goes for TF ratio. Take for example, given the TF reference score as 16.49, a candidate keyphrase which holds 17 as the TF score will be given a higher likelihood rate compares to a candidate keyphrase that holds 30 as the TF score. With this information, the ranking score will give a better result.

In summary, TF, TF ratio, First Occurrence, and Last Occurrence are suitable for keyphrase extraction. Term Cohesion Degree and Phrase Position in Sentences are not reliable and should be used with care.

V. Conclusion and Future Work

This paper sets out to understand the relation between five attribute scores and author-supplied keyphrases. From the results, it can be concluded that three attributes: TF, First Occurrence, and Last Occurrence are good attributes to be retained for weighting and ranking process in automatic keyphrase extraction. Frequency dependent attribute (i.e., Term Cohesion Degree) and Position in Sentences are discarded for the moment as more research need to be done before further conclusion can be made. In addition, TF ratio can be used as a threshold attribute since it gives a consistent range of ratio that author-supplied keyphrases occupy less than 1% of a document.

The results obtained here serve as a first step to improve the accuracy of automatic keyphrase extraction. One limitation of the study is that only five commonly used attributes were being tested. Future research should examine other existing attributes or introduce new attributes that will further improve accuracy rate. Another limitation of this study is that the experiments were performed using one single dataset. Future research should use other datasets to further confirm the results of this study. However, since the dataset adopted here was used in [9], it can be trusted as a good dataset.

References

[1] K. Sarkar, M. Nasipuri, and S. Ghose, "A New Approach to Keyphrase Extraction Using Neural Networks," IJCSI International Journal of Computer Science Issues, vol. 7, no. 2, 2010.

[2] N. Kumar and K. Srinathan, "Automatic keyphrase extraction from scientific documents using N-gram filtration technique," Proceeding of the eighth ACM symposium on Document engineering - DocEng '08, p. 199, 2008.

[3] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the C-value/NC-value method," International Journal on Digital Libraries, vol. 3, no. 2, pp. 115–130, Aug. 2000.

[4] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning, "Domain-Specific Keyphrase Extraction," in Proceedings of the 14th ACM international conference on Information and knowledge management, 2005, pp. 668–671.

[5] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-manning, "KEA: Practical Automatic Keyphrase Extraction," in Proceedings of the fourth ACM conference on Digital libraries, 1999.

[6] P. Turney, "Learning to Extract Keyphrases from Text," National Research Council of Canada, 1999.

[7] A. Csomai, "Keywords in the mist: Automated keyword extraction for very large documents and back of the book indexing.," University Of North Texas, 2008.

[8] P. D. Turney, "Extraction of Keyphrases from Text: Evaluation of Four Algorithms," October, p. 31, 1997.

[9] S. R. El-Beltagy and A. Rafea, "KP-Miner: A keyphrase extraction system for English and Arabic documents," Information Systems, vol. 34, no. 1, pp. 132–144, Mar. 2008.

[10] S. N. Kim and M.-Y. Kan, "Re-examining automatic keyphrase extraction approaches in scientific articles," Proceedings of the Workshop on Multiword Expressions Identification, Interpretation, Disambiguation and Applications - MWE '09, no. August, p. 9, 2009.

[11] O. Medelyan and I. H. Witten, "Domain-Independent Automatic Keyphrase Indexing with Small Training Sets," Journal of the American Society for Information Science & Technology, vol. 59, no. 7, pp. 1026–1040, 2008.

[12] Y. Park, R. J. Byrd, and B. K. Boguraev, "Automatic Glossary Extraction: Beyond Terminology," in Proceedings of the 19th international conference on Computational linguistics - Volume 1, 2002.

[13] P. D. Turney, "Learning Algorithms for Keyphrase Extraction," Information Retrieval - INRT 34-99, 1999.