ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 3, May2012

Text MiningUsing Coherent Keyphrase Extraction

[1#]Prof. Manjusha Yeola, [1#]Prof. Mamta Hegde
[1]Department of Computer Engineering
MIT's MAE, Alandi (D), Pune.
[#]*University of Pune, Maharashtra.*

***Abstract*— A fundamental problem that frequently arises in a great variety of fields such as pattern recognition, image processing, machine learning is the clustering problem. In its basic form the clustering problem is defined as the problem of finding homogeneous groups of data points in a given data set. Each of these groups is called a cluster in which the density of objects is locally higher than in other regions. A popular clustering method that minimizes the clustering error is the k-means algorithm. However, the k-means algorithm is well known that its performance heavily depends on the initial starting conditions. To solve this problem, this paper proposes a new clustering algorithm based on the coherent keyphrase extraction algorithm. In this paper, documents are grouped into several clusters, but the number of clusters is automatically determined by finding out the similarities between documents and the extracted keyphrases**.

***Index Terms*— Keyphrases, cluster, coherent, machine learning, lexical cohesion etc.**

I. INTRODUCTION

The simplest form of clustering is partitional clustering which aims at partitioning a given data set into disjoint subsets (clusters) so that specific clustering criteria are optimized. The most widely used criterion is the clustering error criterion which for each point computes its squared distance from the corresponding cluster center and then takes the sum of these distances for all points in the data set.

A popular clustering method that minimizes the clustering error is the k-means algorithm. However, the k-means algorithm is a local search procedure and it is well known that it suffers from the serious drawback that its performance heavily depends on the initial starting conditions [1].

To solve this problem, this paper proposes a new clustering algorithm based on document's keyphrases which improves the traditional K-means algorithm. The Global k-means means clustering algorithm with the coherent keyphrase extraction algorithm which returns several keyphrases from the source documents by using some machine learning techniques [2].

Keyphrases are useful for a variety of purposes, including summarizing, indexing, labeling, categorizing, clustering, highlighting, browsing, and searching. The task of automatic keyphrase extraction is to select keyphrases from within the text of a given document. Automatic keyphrase extraction makes it feasible to generate keyphrases for the huge number of documents that do not have manually assigned keyphrases [4]. A limitation of previous keyphrase extraction algorithms is that the selected keyphrases are occasionally incoherent. That is, the majority of the output keyphrases may fit together well, but there may be a minority that appears to be outliers, with no clear semantic relation to the majority to each other. This paper presents enhancements to the Kea keyphrase extraction algorithm that are designed to increase the coherence of the extracted keyphrases. The approach is to use the degree of statistical association among candidate keyphrases as evidence that they may be semantically related [2].

II. RELATED WORK

A. Coherent Keyphrase Extraction Algorithm

A limitation of prior keyphrase extraction algorithms is that the output keyphrases are at times incoherent. Discarding the incoherent candidates might improve the quality of the machine-extracted keyphrases. The approach is to measure the degree of statistical association among the candidate phrases. The hypothesis is that semantically related phrases will tend to be statistically associated with each other, and that avoiding unrelated phrases will tend to improve the quality of the output keyphrases.

Recent work on text summarization has used lexical cohesion in an effort to improve the coherence of machine generated summaries. Instead of using a thesaurus, if we use

statistical word association to estimate lexical cohesion it can improve the performance.

In the simplest version of Kea, candidate phrases are classified using only two features: *TFxIDF* and *distance*. This paper introduces a new set of features for measuring coherence. After training, given a new input document and a desired number of output phrases, *N*, Kea converts the input document into a set of candidate phrases with associated feature vectors. It uses the naive Bayes model to calculate the probability that the candidates belong to the class *keyphrase*, and then it outputs the *N* candidates with the highest probabilities.

1. Coherence Feature Set

The coherence feature set is calculated using a two-pass method. The first pass processes the candidate phrases using the baseline feature set. The second pass uses the top *K* most probable phrases, according to the probability estimates from the first pass, as a standard for evaluating the top *L* most probable phrases *(K< L).* In the second pass, for each of the top *L* candidates (including the top *K* candidates), new features are calculated based on the statistical association between the given candidate phrase and the top *K* phrases.

The hypothesis is that candidates that are semantically related to one or more of the top *K* phrases will tend to be more coherent, higher quality keyphrases.

Experiments done by scinetists show that the coherence features are a significant improvement over the baseline features alone. More of the output keyphrases match with the authors' keyphrases, which is evidence that their quality has improved [2].

B.Global K-means Algorithm

Global k-means algorithm which is an incremental approach to clustering that dynamically adds one cluster center at a time through a deterministic global search procedure consisting of N (with N being the size of the data set) executions of the k-means algorithm from suitable initial positions.

The algorithm proceeds in an incremental way: to solve a clustering problem with M clusters, all intermediate problems with 1; 2; : : : ; M −1 clusters are sequentially solved. The basic idea underlying the proposed method is that an optimal solution for a clustering problem with M clusters can be obtained using a series of local searches (using the k-means algorithm). At each local search the M − 1 cluster centers are always initially placed at their optimal positions corresponding to the clustering problem with M − 1 clusters. The remaining M^{th} cluster center is initially placed at several positions within the data space. Since for M =1 the optimal solution is known, we can iteratively apply the above procedure to 2nd optimal solutions for all k-clustering problems k =1; : : : ; M. In addition to effectiveness, the method is deterministic and does not depend on any initial conditions or empirically adjustable parameters.

These are significant advantages over all clustering approaches [1].

C. Global k-means keyphrase Extraction Algorithm

This is our proposed new clustering method that can improve the Global K-means algorithm by combining it with the coherent keyphrase extraction algorithm.

In our proposed algorithm documents are clustered into several groups using Global K-means with coherent keyphrase extraction algorithm which determines candidate keyphrases that are semantically related to one other and out of that will find the top *K* phrases tend to be more coherent, higher quality keyphrases.

Global k-means clustering algorithm constitutes a deterministic global optimization method that does not depend on any initial parameter values and employs the k-means algorithm as a local search procedure. Instead of randomly selecting initial values for all cluster centers as is the case with most global clustering algorithms the proposed technique proceeds in an incremental way attempting to optimally add one new cluster center at each stage.

More specifically, to solve a clustering problem with M clusters the method proceeds as follows. We start with one cluster (k = 1) and find its optimal position which corresponds to the centroid of the data set X. In order to solve the problem with two clusters (k =2) we perform N executions of the k-means algorithm from the following initial positions of the cluster centers: the first cluster center is always placed at the optimal position for the problem with k = 1, while the second center at execution n is placed at the

position of the data point X_n (n=1; : : : ; N). The best solution obtained after the N executions of the k-means algorithm is considered as the solution for the clustering problem with k =2. In general, let (m_1*(k); : : : ; m_k* (k)) denote the final solution for k-clustering problem. Once we have found the solution for the (k − 1)-clustering problem, we try to find the solution of the k-clustering problem as follows: we perform N runs of the k-means algorithm with k clusters where each run n starts from the initial state (m_1* (k−1); : : : ; $m_{(k-1)}$*(k−1); xn).

The best solution obtained from the N runs is considered as the solution (m_1* (k); : : : ; m_k* (k)) of the k-clustering problem.

By proceeding in the above fashion we finally obtain a solution with M clusters having also found solutions for all k-clustering problems with k <M [1].

The coherence feature set is calculated using a two-pass method. The first pass processes the candidate phrases using the baseline feature set. The second pass uses the top *K* most probable phrases, according to the probability estimates from the first pass, as a standard for evaluating the top *L* most probable phrases *(K< L).* In the second pass, for each of the top *L* candidates (including the top *K* candidates), new features are calculated based on the statistical association between the given candidate phrase and the top *K* phrases.

The hypothesis is that candidates that are semantically related to one or more of the top *K* phrases will tend to be more coherent, higher quality keyphrases [2].

These keyphases are used to assign weights to other phrases. Now, the distance between the weighted documents and centroids are measured, and if the measured values do not reach to the threshold value, the value of *k* is increased by 1. This process is repeated until the measured distance exceeds the threshold value. At this point, the number of clusters *k* is determined, and the Global K-means algorithm now can be used for actual clustering [3].

III. CONCLUSION

A limitation of previous keyphrase extraction algorithms is that the selected keyphrases are occasionally incoherent. That is, the majority of the output keyphrases may fit together well, but there may be a minority that appears to be outliers, with no clear semantic relation to the majority or to each other. Coherence kayphrase extraction algorithm enhances the Kea keyphrase extraction algorithm designed to increase the coherence of the extracted keyphrases [2].

Also the global k-means clustering algorithm, which constitutes a deterministic effective global clustering algorithm for the minimization of the clustering error that, employs the k-means algorithm as a local searchprocedure.

This method can provide easy and efficient ways to extract test documents from massive quantity of resources.

The main limitation of the new coherence feature set is the time required to calculate the features

REFERENCES

[1]. Likas, N. Vlassis, J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451-461, February 2003.

[2]. P. Turney. Coherent keyphrase extraction via web mining. *Technical Report ERB-1057, Institute for Information Technology, National Research Council of Canada*, 1999.

[3]. Juhyun Han, Taehwan Kim, Joongmin Choi, Web Document Clustering By Using Automatic Keyphrase Extraction *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops* 2007.

[4]. I. Witten, G. Paynter, E. Frank, C. Gutwin, C. Nevill-Manning. Kea: Practical automatic keyphrase extraction. *Proc. 4th ACM Conference on Digital Libraries*, 254-255, August 1999.

[5]. M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques. *Proc. KDD Workshop on Text Mining*, 1-20, 2000.

[6]. J. Wu, A. Agogino. Automating keyphrase extraction with multi-objective genetic algorithms. *Proc. 37th Annual Hawaii International Conference on System Sciences (HICSS)*, 104-111, 2004.

[7]. P. Turney. Learning to extract keyphrases from text. *Proc. 18th International Joint Conference on Artificial Intelligence (IJCAI)*, 434-439, 2003.

[8]. J.A. Lozano, J.M. Pena, P. Larranaga, An empirical comparison of four initialization methods for the k-means algorithm, Pattern Recognition Lett. 20 (1999) 1027–1040.

[9]. Jones, S. and Paynter, G.W. Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications. *Journal of the American Society for Information Science and Technology (JASIST)*, 53 (8), 653-677, 2002.

[10]. Turney, P.D. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2, 303-336, 2000.