# Multi-Document Text Summarization Using Random Indexing and Clustering

Ambrish Rawat[a,1], Pramod Kumar Sahoo[b], Niladri Chatterjee[c]

[a, c] *Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, India*
[b] *Institute for Systems Studies & Analyses, Defence Research & Development Organisation, Delhi, India*

## Abstract

Extractive text summarization for a set of documents suffers from two serious problems, viz. repetition of information in several documents, complementary information existing on the same theme across different documents. As a consequence, algorithms applicable for single document extractive summarization do not work well in multi-document scenario. The present paper proposes a clustering based technique where the sentences of the document collection are clustered based on their themes. The sentences in each cluster are then ranked according to their importance. The ranked sentences are then picked in a systematic manner to generate the summary. Random Indexing, which has assumed significant attention in recent times, has been used as the representation scheme for the documents in order to avoid large binary vectors as is the case with typical Word Space Model. The system has been tested with DUC 2002 and DUC 2007 datasets. The proposed algorithm is found to perform better than some existing algorithms as per the scores given by various ROUGE metrics.

*Keywords:* Multi-Document Text summarization; Random Indexing; Clustering; Sentence Extraction

## 1. Introduction

In extractive text summarization representative sentences are chosen from a document for generation of the summary [Mani, 2001]. One of the ways of achieving this is to represent sentences in a multi-dimensional word space. Word Space Model (WSM) is a technique where vector representation of sentences is used for representing a given text in a Word Space [Schütze, 1993]. Typically, these vectors are of very large dimension. Hence an efficient representation of text using WSM is computationally intensive for a large set of documents. One of the most common WSM-based approach is Latent Semantic Analysis (LSA) [Landauer & Dumais, 1997; Landauer *et al*., 1998]), where orthogonal unary vectors are used to represent the words of a document. The dimension of these vectors is equal to the number of unique words in the document. For each sentence of the document its *sentence vector* is then computed as weighted sums of the word vectors corresponding to the constituent words of the underlying sentence. However, large dimension of the vectors make them computationally challenging. Although dimension reduction techniques, such as Singular Value Decomposition (SVD) [Klema & Laub, 1980], is often used for dimensionality reduction, it renders the whole procedure computationally more expensive.

Random Indexing [Sahlgren, 2005] provides a scalable alternative to LSA where near-orthogonal index vectors are used for achieving an incremental learning of the context.

[1] Present address: Department of Engineering, University of Cambridge, Cambridge, UK

Moreover, by fixing a lower dimension for the index vectors used for word representation, one avoids the computationally expensive task of SVD. Random Indexing has already been used for extractive text summarization of single documents by Chatterjee & Mohan (2007) and Gustavsson & Jönsson (2010). An improved formulation of Random Indexing based single document extractive summarization has also been suggested in [Chatterjee & Sahoo, 2013]. A straightforward application of single document summarization strategies in a multi-document scenario, however, suffers from some serious problems. Some of them are:

- The size of the document collection may be very large in comparison with a single document. For instance, one particular collection in the DUC[2] 2007 dataset comprised of 24 documents with roughly 22000 words making the summary generation a challenging task.
- A set of documents may have a common theme. For instance, in the collection D0701 of DUC 2007 dataset, many documents have a sentence which in some form conveys the information that "Morris Dees is the co-founder of the Southern Poverty Law Centre" (see Table 1). A summarizer for this collection is not expected to capture these different sentences in order to avoid redundant information in the generated summary. A multi-document summarizer should therefore have the ability to minimize this.

Table 1. Excerpts from three documents of D0701 collection from DUC 2007 Dataset

| Document Number | Text on Morris Dees |
|---|---|
| **APW20000909.0144** | Morris Dees, co-founder of the Southern Poverty Law Center in Montgomery, Ala., represented the Keenans and has said he intends to take everything the Aryan Nations owns to pay the judgement, including the sect's name. |
| **APW20000908.0184** | Morris Dees, the co-founder of the Southern Poverty Law Center in Montgomery, Ala., and one of the attorneys for the plaintiffs, said he intended to enforce the judgement, taking everything the Aryan Nations owns, including its trademark name. |
| **NYT19980715.0137** | In the same issue, Morris Dees, leader of the Southern Poverty Law Center, gets an unwarranted slap in the Media Watch column. Dees has always gotten props for dropping dimes on the hateful activities of such groups as the Ku Klux Klan and Aryan Nation. |

- It is also important that a multi-document text summarizer incorporates different information about a common prevalent theme spread across different documents. For example, many documents in D0701 describe some piece of information about the "Southern Poverty Law Center" in different contexts (see Table 2). A summarizer for

[2] http://duc.nist.gov/ (Accessed on 11th Dec 2015)

this collection should be able to capture this as a theme, and include different sentences pertaining to this theme but focussing on different aspects of the same.

- A significant difference between single-document and multi-document extractive text summarization stems from the redundancy of the chosen sentences, and also the compression ratio of the summary. Since the degree of redundancy among the sentences from a set of related documents is much higher than among the sentences from a single document, it is essential that a multi-document text summarizer adopts measures to address redundancy.

Table 2. Excerpts from three documents of D0701 collection from DUC 2007 Dataset

| Document Number | Text on Southern Poverty Law Center |
|---|---|
| **NYT19990302.0069** | But the book didn't begin, nor will it end with the King trial, as a report by the Montgomery, Ala.-based Southern Poverty Law Center demonstrates. According to the report, the number of hate groups in the country increased again last year. |
| **NYT19990304.0736** | Since co-founding the Southern Poverty Law Center in 1971, Dees has wielded the civil lawsuit like a buck knife, carving financial assets out of hate group leaders who inspire followers to beat, burn and kill. |
| **NYT19990822.0227** | Lawyers from the Southern Poverty Law Center, a civil rights organization in Montgomery, Ala., advanced his case in federal court, charging that the Escude Funeral Home and Hixson Brothers Funeral Home in Avoyelles Parish either refused to deal with blacks or offered "distinctly inferior services" for the same prices that they charged whites. |

Two generic approaches in this regard can be found in literature:

- By measuring the similarity of new candidate sentences with respect to the previously selected ones, and pick the one that is least similar to the previously selected ones [Carbonell & Goldstein, 1998].
- Another approach is by clustering the sentences from the document collection and thereby capturing the multiple themes present in the collection [Radev *et al.*, 2004]. A summary can thereafter be generated by selecting representative sentences from different clusters.

Over the last decade or so several multi-document summarization approaches based on WSM, Natural Language Processing (NLP) techniques, and clustering methods have been proposed in literature. Wang *et al*. (2008) proposed a multi-document summarization framework based on sentence-level semantic analysis (SLSS) and symmetric non-negative matrix factorization (SNMF). The experiments on DUC 2005 and DUC 2006 datasets shows the effectiveness of their method over LSA [Gong & Liu, 2001] and NMFBase [Lee & Seung, 2001] methods.

Hong & Nenkova (2014) proposed RegSum system which employs a supervised model for predicting word importance. RegSum combines the weights estimated from three unsupervised approaches, along with features including locations, part-of-speech, name-entity-tags, topic categories and contexts. Specifically, this system captures words which are of intrinsic interest to people by analyzing a large number of summary-abstract pairs from the New York Times corpus [Sandhaus, 2008]. RegSum trained on DUC 2003 dataset and evaluated on DUC 2004 dataset reveals that the quality of the summaries could be greatly improved by better estimation of word importance.

Several clustering-based multi-document summarization methods have been proposed in literature. Wan & Yang (2008) proposed two models ClusterCMRW and ClusterHITS to incorporate the cluster-level information into the process of sentence ranking. ClusterCMRW incorporates the cluster-level information into the link graph using conditional Markov random walk model. ClusterHITS considers the clusters and sentences as hubs and authorities in the HITS algorithm. These two models found to outperform the top performing participant systems of DUC 2001 and DUC 2002.

Cao *et al*. (2015) proposed a Ranking framework upon Recursive Neural Networks (R2N2) for the sentence ranking task of multi-document summarization. The performance of R2N2 found to be very close to ClusterHITs (DUC 2001 dataset), ClusterCMRW (DUC 2002 dataset), and comparable to RegSum (DUC 2004 dataset).

Sun *et al*. (2009) proposed RankClus - a clustering framework to integrate clustering and ranking. RankClus generates conditional ranking relative to clusters to improve ranking quality. Further, it uses conditional ranking to generate new measure attributes to improve clustering. As a result, the quality of clustering and ranking are mutually enhanced, which means the clusters are getting more accurate and the ranking is getting more meaningful.

Cai & Li (2011) proposed a spectral analysis based multi-document summarization approach that simultaneously creates clusters and rank sentences by investigating the spectral characteristics of the similarity network which is constructed upon the document(s). To improve upon spectral analysis based approach Cai & Li (2013) proposed another clustering technique applied on bi-type document graph. In this approach three different ranking functions, viz. global ranking, local ranking and conditional ranking, have been defined in a bi-type document graph constructed from the given document set. Based on initial clusters, ranking is applied separately on the document graph. Then a mixture model is applied to

decompose each sentence into a multi-dimensional vector, where each dimension is a component coefficient with respect to a cluster, which is measured by rank distribution. Then the sentences are reassigned to the nearest cluster under the new measure space to improve clustering. As a result, quality of clustering and ranking are mutually enhanced. Experimental results on the DUC 2004–2007 datasets demonstrates the effectiveness of the approach proposed by Cai & Li (2013) over ClusterHITS, RankClus, and Spectral Analysis based approaches.

The present paper proposes a hybrid approach of the Random Indexing based semantic representation along with traditional clustering techniques to achieve a summary for a given set of documents. In Section 2 we describe the general principles of Random Indexing. Section 3 discusses the proposed strategy which is a combination of the Random Indexing based representation proposed for single-document summarizer and the clustering algorithms. The experimental results have been summarized in Section 4 and the conclusion and future work have been briefed in Section 5.

## 2. Random Indexing

Random Indexing aims at assigning near-orthogonal index vectors of fixed dimension to a set of words for a smaller and denser representation in the Word Space. These index vectors are randomly assigned in such a manner that they are unique (i.e. each distinct word gets a different index vector), sparse and ternary. Every vector consists of values 0, +1 and -1, which are randomly assigned according to the following distribution:

$$\begin{cases} +1 & \text{with probability} \quad \dfrac{n_\varepsilon}{2d_I} \\ 0 & \text{with probability} \quad \dfrac{d_I - n_\varepsilon}{d_I} \\ -1 & \text{with probability} \quad \dfrac{n_\varepsilon}{2d_I} \end{cases} \tag{1}$$

Here, $d_I$ is the fixed dimension of the index vectors and $n_\varepsilon \ll d_I$. We have conducted experiments with different values of $n_\varepsilon$ for random indexing based single document text summarization, where the value of $d_I$ depends on the size of the document. In the present work, we have experimented with three different values of $n_\varepsilon$ viz. 2, 4 and 6, and correspondingly three different values of $d_I$ as suggested in [Sahoo, 2014]. With an assumption that the context of a word is determined by its neighbouring words, the context

vectors for all the words are calculated by adding the index vectors of neighbouring words. In our experimental framework, we have defined the neighbourhood of a word as consisting of the two words succeeding and the two words preceding the word under consideration.

## 3. Clustering based multi-document text summarization

The proposed strategy for text summarization consists of the following three steps:

- Representation of the sentences in a document collection. After creating the word level index vectors as described in Section 2, the document sentences are represented in the form of sentence vectors in the Word Space through aggregation of the constituent index vectors.
- Clustering of the sentence semantic vectors in the proximity graph. After mapping the sentences into the Word Space, the entire document is converted into a proximity graph. In this graph the sentences are represented as nodes, and edge weights between two sentences give the strength of semantic proximity between sentences.
- Selection of representative sentence(s) from the ranked clusters for summary generation.

In this work, we have experimented with different index vector assignments along two different clustering strategies (explained in Section 3.2), and two different sentence selection strategies (explained in Section 3.3) for subsequent summary generation. We also experimented with three different distance measures in the clustering and sentence selection strategies.

### *3.1. Representation of sentences in the Word Space*

Semantic vector representation of the sentences in a document collection has been achieved in four steps:

- Preprocessing of the available textual data from the collection;
- Generation of index vectors for the unique words;
- Computation of context vectors for each unique word in the collection;
- Generation of sentence semantic vectors for each sentence in the document collection.

### *3.1.2. Preprocessing*

As a first step, the texts from all the documents in a document collection are combined into one single text file. This is followed by a sequence of text-processing steps: removal of common words and stemming. A set of commonly occurring words is extracted from the combined document to generate a list of content words. Then, common morphological inflectional endings are removed from these content words using Porter Stemming Algorithm [Porter, 1980], and an updated list of unique content words $\{u_i\}$ is obtained.

### *3.1.3. Generation of Index Vectors*

This is done as per the scheme explained in Section 2. A unique ternary index vector $I(u_i)$ is generated for each word $u_i$ in the unique content words list for a given value of $n_\varepsilon$. Here, $n_\varepsilon$ is defined as the number of '+1's and '-1's assigned to every index vector. In this work we have experimented and compared results for three different values of $n_\varepsilon$, viz. 2, 4 and 6. As per suggestions in [Chatterjee & Sahoo, 2015] we have restricted the placement of the `+1's to the upper half of an index vector, and the placement of the `-1's to the lower half. The dimension of the index vector $d_I$ has been varied according to both $n_\varepsilon$ and $m$, where $m$ is the total number of unique content words in the document collection [Chatterjee & Sahoo, 2015].

$$d_I = \begin{cases} 2\lceil m \rceil & \text{if} \quad n_\varepsilon = 2 \\ 2\left\lceil 0.5\left(1+\sqrt{1+8\sqrt{m}}\right)\right\rceil & \text{if} \quad n_\varepsilon = 4 \\ 2x : x \in \mathbb{R} \wedge x^3 - 3x^2 + 2x - 6\sqrt{m} = 0 & \text{if} \quad n_\varepsilon = 6 \end{cases} \tag{2}$$

Here, $\lceil\ \rceil$ denotes the ceiling function, and $d_I$ denotes the minimum dimension required for generating random indexing based index vectors for all the distinct words of a document.

### *3.1.4. Generation of Context Vectors*

A context vector for each word in the unique content words list is iteratively generated be adding index vectors of its neighbourhood. As mentioned in Section 2, we have considered the two succeeding and the two preceding words along with the word under consideration, for the neighbourhood definition. This gives us the following formulation for the context vector $C(w_{i,j})$ for the word $w_{i,j}$ occurring at $i^{\text{th}}$ position of the $j^{\text{th}}$ sentence of the combined document:

$$C(w_{i,j}) \coloneqq C(w_{i,j}) + \sum_{\substack{k=-2 \\ k\neq 0}}^{2} 2^{1-|k|}\, I(w_{i+k,j}) \tag{3}$$

where $I(w_{i,j})$ is the index vector of the word $w_{i,j}$ and $2^{1-|k|}$ is the weighting factor of the neighbouring word which accounts for its proximity to the word under consideration.

*3.1.5. Generation of Sentence Semantic Vectors*

The sentence semantic vector, denoted as $S(s_j)$, for the $j^{\text{th}}$ sentence $s_j$ of the combined document is defined as:

$$S(s_j) = \frac{1}{m_j} \sum_{i=1}^{m_j} \big(C(w_{i,j}) - T\big) \tag{4}$$

where $m_j$ is the number of content words in $s_j$, and $T$ denotes the central theme of the document. The central theme of the document collection is computed using the arithmetic mean of the context vectors of all the content words in the given document collection. Hence, $T$ is computed as:

$$T = \frac{1}{\sum_{j=1}^{m} m_j} \sum_{j=1}^{m} \sum_{i=1}^{m_j} C(w_{i,j}) \tag{5}$$

where $m$ is the number of sentences in the whole document. Subtraction of central theme from the context vectors takes care of the commonly occurring words which unlike in context vector computation are not ignored here [Higgins & Burstein, 2007]. The sentence semantic vectors thus computed provide a geometrical representation of the semantic being conveyed by the sentence. A sentence semantic vector in itself does not convey much information. It merely represents the location of the sentences in the Word Space. It is the relative distance between the sentences that captures the similarity in the meaning of sentence. Two sentences (as represented by their semantic vectors) that lie close to each other in the Word Space have similar meaning.

### *3.2. Clustering the sentences in the proximity graph*

Once each sentence in the entire document collection (to be called $\mathbb{D}$, henceforth) is assigned a semantic vector in the Word Space, we represent $\mathbb{D}$ by proximity graph. The nodes

of the graph represent the sentences of $\mathbb{D}$. The distance between two nodes is representative of the similarity in theme and meanings of sentences being represented by those nodes. We exploit this relative thematic similarity of closely spaced sentences in the clustering strategies. These strategies capture the multiple themes present in $\mathbb{D}$ as a set of clusters, $N = \{N_1, N_2, \cdots, N_{n_c}\}$, where $n_c$ equals the total number of clusters. In this work we have used DUC 2002 and DUC 2007 datasets; and experimented with four different values of $n_c$ namely, $n_c = 5,10,15,20$.

We employed two traditional unsupervised clustering techniques, viz. K-Means Clustering [Hartigan & Wong, 1979] and Clustering based on Expectation Maximization of Gaussian Mixture Models (EMGM-Clustering) [Redner & Walker, 1984]. Each of these strategies utilises different mathematical framework while clustering the data points. We have used three different distance measures namely, Euclidean distance $(D_E)$, cosine dissimilarity $(D_C)$ and angular distance $(D_A)$ - for calculating the distance between any pair of sentence semantic vectors, $S(s_i) = [s_{i1}, s_{i2}, \cdots, s_{id_I}]$ and $S(s_j) = [s_{j1}, s_{j2}, \cdots, s_{jd_I}]$, in the Word Space. Equations (6), (7) and (8) give the mathematical formulation of the three distance metrics.

$$D_E\left(S(s_i), S(s_j)\right) = \sqrt{\sum_{t=1}^{d_I} \left(s_{it} - s_{jt}\right)^2} \tag{6}$$

$$D_C\left(S(s_i), S(s_j)\right) = 1 - \frac{\sum_{t=1}^{d_I} \left(s_{it} \cdot s_{jt}\right)}{\sqrt{\sum_{t=1}^{d_I} s_{it}^2} \cdot \sqrt{\sum_{t=1}^{d_I} s_{jt}^2}} \tag{7}$$

$$D_A\left(S(s_i), S(s_j)\right) = \frac{1}{\pi} \cos^{-1}\left(\frac{\sum_{t=1}^{d_I} \left(s_{it} \cdot s_{jt}\right)}{\sqrt{\sum_{t=1}^{d_I} s_{it}^2} \cdot \sqrt{\sum_{t=1}^{d_I} s_{jt}^2}}\right) \tag{8}$$

In the subsequent discussions a distance metric in general will be denoted as $D$, unless it is specified otherwise.

### *3.2.1. K-Means Clustering*

In this strategy the association of a sentence semantic vector $S(s_i)$ with a cluster is computed iteratively such that the inter-cluster distance between a pair of sentence semantic vectors belonging to different clusters is maximised; while the intra-cluster distance between a pair of sentence semantic vectors belonging to the same cluster is minimised [Hartigan &

Wong, 1979]. In every iteration the sentence semantic vector $S(s_i)$ is assigned to a cluster $N_k$ if Equation (9) holds good:

$$D\big(m_k, S(s_i)\big) = \min_{p=1,2,\cdots,n_c} D\left(m_p, S(s_i)\right) \tag{9}$$

where $m_p$ is the cluster centroid of the $p^{\text{th}}$ cluster $N_p$ which is defined as the arithmetic mean of all sentence semantic vector belonging to that cluster. Thus, on convergence, a pair of sentences $s_i$ and $s_j$ with similar themes will have their sentence semantic vectors $S(s_i)$ and $S\big(s_j\big)$ in the same cluster. But, a pair of sentences $s_a$ and $s_b$ with different themes will have their sentence semantic vectors $S(s_a)$ and $S(s_b)$ in different clusters.

### *3.2.2. EMGM Clustering*

EMGM clustering uses Expectation Maximization algorithm for fitting the Gaussian Mixture Models (GMM) to cluster the sentence semantic vectors [Redner & Walker, 1984]. It is assumed that the set of sentence semantic vectors have a joint distribution given by $p\big(S(s_i)\big)$ with $n_c$ Gaussian components. Thus, it is assumed that a cluster $N_k$ has a multivariate Gaussian distribution $p_k(\mu_k, \Sigma_k)$ in the Word Space with $\mu_k$ and $\Sigma_k$ as the mean vector and the covariance matrix respectively of the underlying Gaussian distribution:

$$p(S(s_i)|\Theta) = \sum_{k=1}^{n_c} \alpha_k p_k(S(s_i)|\mu_k, \Sigma_k) \tag{10}$$

where,

$$p_k(S(s_i)|\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^{d_I}|\Sigma_k|}} exp\left(-\frac{1}{2}(S(s_i) - \mu_k)^T \Sigma_k^{-1}(S(s_i) - \mu_k)\right) \tag{11}$$

In the above expressions $\Theta$ is the set of parameters $\alpha_k$, $\mu_k$ and $\Sigma_k$. The parameter $\alpha_k \geq 0$ for each $k$, and $\sum_{k=1}^{n_c} \alpha_k = 1$. Since $\Theta$ is unknown they need to be estimated. First we initialize them randomly, and then update them iteratively so as to maximize the likelihood of assignment of semantic vectors to a Gaussian sample.

Each cluster is considered to be a representative of a theme from the document collection. The theme of the cluster $N_k$ is denoted by $T_k$, which is calculated as the arithmetic mean of

the semantic vectors of all the sentences in that cluster. If $b_k$ denotes the total number of sentence semantic vectors in $N_k$ cluster, and $s_i^k$ denotes the $i^{\text{th}}$ sentence in $N_k$, then we have:

$$T_k = \frac{1}{b_k} \sum_{i=1}^{b_k} S(s_i^k) \tag{12}$$

Extending the reasoning from the previous section, the relative distance between cluster themes exemplifies the similarity in themes of the clusters. This gives a natural ranking to the clusters according to their distances from the central theme of the document collection. Each cluster $N_k$ is thus given a rank $R(N_k) = r_k$ for $r_k \in \{1,2,\cdots,n_c\}$.

### *3.3. Representative sentence selection from the ranked clusters and summary generation*

In order to obtain candidate sentences for a summary, sentences are selected in a defined order from the formed clusters. It was noted that selecting multiple sentences from one cluster will lead to redundant sentences in a summary which convey similar meaning pertaining to that particular cluster's theme. Also, selection of sentences from a cluster that is not close to the central theme in the Word Space, will lead to less relevant sentences in a summary. In order to balance the relevance and redundancy of sentences in a summary, we devised the following scheme.

Starting from cluster $N_p$ with $R(N_p) = 1$, we chose the $1^{\text{st}}$ sentence which is the best (i.e. sentence rank = 1) representative of the cluster theme of that document. Similarly, the $2^{\text{nd}}$ sentence in the summary is the best (sentence rank = 1) representative of the cluster $N_q$, where $R(N_q) = 2$. The process is continued till best (sentence rank = 1) representative from all the $n_c$ clusters has been selected. This is followed by a selection of the $2^{\text{nd}}$ best (i.e. sentence rank = 2) representative from all $n_c$ clusters in the order of their rank and so on. The $n^{\text{th}}$ best representative sentence of a cluster $N_p$ has been computed using the following two strategies:

1. The $n^{th}$*closest to the centroid*: In this strategy, we arrange the sentences in a cluster $N_p$ according to their distance $D$ from the cluster theme $T_p$. The candidacy of a sentence in a cluster is thus determined by its proximity to the corresponding cluster's theme. This is in accordance with the reasoning that relative distance between a pair of semantic vectors captures the similarity in meanings of the two vectors. This in turn implies that the closer a sentence is to its cluster's theme, the better is its relevance in conveying the theme of that

cluster. This results in the following relationship between the $i^{\text{th}}$ sentence $s_i^p$ and $j^{\text{th}}$ sentence $s_j^p$ of the cluster $N_p$:

$$D\left(S\left(s_i^p\right), T_p\right) \leq D\left(S\left(s_j^p\right), T_p\right) \Longrightarrow R\left(s_i^p\right) \geq R\left(s_j^p\right) \tag{13}$$

2. The $n^{th}$ *longest sentence*: In this strategy, we arrange the sentences according to their length. The length of the $i^{\text{th}}$ sentence $s_i^p$ of a cluster $N_p$ denoted as $L\left(s_i^p\right)$ is defined as the number of content words in that sentence $s_i^p$. Under this strategy the sentence with the maximum length in a cluster is considered as the *best* (sentence rank = 1) representative of the cluster. This means,

$$L\left(s_i^p\right) \leq L\left(s_j^p\right) \Longrightarrow R\left(s_i^p\right) \leq R\left(s_j^p\right) \tag{14}$$

Selecting the representative sentences iteratively using either of the two strategies described above, the summary of the document collection is generated till summary of the desired length is obtained.

## 4. Experimental Results

We evaluated our summarization scheme on DUC 2002 and DUC 2007 datasets. DUC 2002 dataset consists of 567 documents divided into 59 document sets (D088 is excluded from the original 60 document sets by NIST) with approximately 10 documents in each document set. For each document set, we generated summaries of 200 word length. DUC 2007 dataset consists of 45 document sets with 25 documents in each document set. For each document set, we generated summaries of 250 word length.

We evaluated the following ROUGE [Lin, 2004] scores for our proposed strategies:

- ROUGE-1: Unigram recall measure
- ROUGE-2: Bigram recall measure
- ROUGE-L: Longest common subsequence recall measure
- ROUGE-SU4: Skip bigram with unigram recall measure (maximum skip distance 4)

As explained in Section 3, five parameters were varied during the experiment:

- $n_\varepsilon$ i.e., the total number of '+1's and '-1's in the index vector. In our experiments $n_\varepsilon \in \{2,4,6\}$.
- $n_c$ i.e., the number of clusters. In our experiment $n_c \in \{5,10,15,20\}$.
- Choice of clustering strategy: K-means clustering and EMGM clustering
- Choice of distance measure: Euclidean distance, Cosine dissimilarity, and Angular distance
- Choice of sentence selection scheme: closest to centroid and longest sentence

This resulted in 144 different summaries for each document collection. The patterns observed from these 144 summaries are analysed in the following sections.

### *4.1. Different values of $n_\varepsilon$*

The three different values of $n_\varepsilon$ resulted in three different dimensions for the index vectors. This led to three different sets of sentence semantic vectors for each document collection. Table 3 gives the ROUGE scores for the summaries obtained with these three sets when 5 clusters were formed using the EMGM clustering strategy and representative sentence from each cluster was selected using closest to centroid scheme. Euclidean distance measure was used in this part of the experiment. It was observed that $n_\varepsilon = 2$ produced the best summaries in our proposed scheme.

Table 3: ROUGE recall scores of summaries generated for different values of $n_\varepsilon$ ($n_c = 5$, EMGM clustering, Euclidean distance measure, Closest to centroid sentence selection scheme)

| **ROUGE Measure** | **DUC 2002 Dataset** | | | **DUC 2007 Dataset** | | |
|---|---|---|---|---|---|---|
| | $\boldsymbol{n_\varepsilon = 2}$ | $\boldsymbol{n_\varepsilon = 4}$ | $\boldsymbol{n_\varepsilon = 6}$ | $\boldsymbol{n_\varepsilon = 2}$ | $\boldsymbol{n_\varepsilon = 4}$ | $\boldsymbol{n_\varepsilon = 6}$ |
| ROUGE-1 | 0.38510 | 0.37100 | 0.36049 | 0.36866 | 0.34896 | 0.34834 |
| ROUGE-2 | 0.13107 | 0.11239 | 0.10978 | 0.06520 | 0.05823 | 0.05991 |
| ROUGE-L | 0.36645 | 0.35415 | 0.34237 | 0.33543 | 0.32165 | 0.32150 |
| ROUGE-SU4 | 0.17396 | 0.15697 | 0.15354 | 0.12266 | 0.13387 | 0.11298 |

### *4.2. Different number of clusters*

The number of clusters were varied with an intention to optimally capture the number of themes present in the document. In most cases the difference in number of clusters did not significantly reflect on the ROUGE scores. Table 4 summarizes one set of results for this experiment.

Table 4: ROUGE recall scores of summaries generated for different numbers of clusters ($n_\varepsilon = 4$, EMGM clustering, Euclidean distance measure, Closest to centroid sentence selection scheme)

| ROUGE Measure | DUC 2002 Dataset | | | | DUC 2007 Dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | $n_c = 5$ | $n_c = 10$ | $n_c = 15$ | $n_c = 20$ | $n_c = 5$ | $n_c = 10$ | $n_c = 15$ | $n_c = 20$ |
| ROUGE-1 | 0.37100 | 0.37096 | 0.36760 | 0.37376 | 0.34896 | 0.36781 | 0.37428 | 0.36833 |
| ROUGE-2 | 0.11239 | 0.11118 | 0.11153 | 0.11663 | 0.05823 | 0.06475 | 0.06575 | 0.06278 |
| ROUGE-L | 0.35415 | 0.35158 | 0.34833 | 0.35528 | 0.32165 | 0.33745 | 0.34133 | 0.33644 |
| ROUGE-SU4 | 0.15697 | 0.15658 | 0.15652 | 0.16077 | 0.11387 | 0.12288 | 0.12468 | 0.12123 |

### *4.3. Different clustering strategies*

As mentioned in Section 3 we experimented with two different clustering strategies: K-means clustering and EMGM clustering. Each of these strategies tried to capture the themes in the document collection in different mathematical frameworks. For most combinations K-means clustering strategy generated summaries with better ROUGE scores. Table 5 summarizes the results for one combination of parameters. In this part of the experiment, Euclidean distance was used as the distance measured.

Table 5: ROUGE recall scores of summaries generated for using different clustering strategies ($n_\varepsilon = 2$, $n_c = 15$, Euclidean distance measure, and Closest to centroid sentence selection scheme)

| ROUGE Measure | DUC 2002 Dataset | | DUC 2007 Dataset | |
|---|---|---|---|---|
| | K-Means | EMGM | K-Means | EMGM |
| ROUGE-1 | 0.38990 | 0.38163 | 0.38312 | 0.36975 |
| ROUGE-2 | 0.13472 | 0.12231 | 0.07023 | 0.06602 |
| ROUGE-L | 0.37033 | 0.36244 | 0.34826 | 0.33678 |
| ROUGE-SU4 | 0.17816 | 0.16689 | 0.13064 | 0.12464 |

### *4.4. Different distance measures*

We experimented with three different distance measures: Euclidean distance, cosine dissimilarity and angular distance. Our purpose is to capture different types of similarity relationships amongst the sentence semantic vectors clustering strategies. The results of this experiment for summaries generated using K-means clustering strategy and closest to centroid

sentence selection strategy have been summarized in Table 6. It has been observed that Euclidean distance measure outperforms other distance measures.

Table 6: ROUGE recall scores of summaries generated for using different distance measures ($n_\varepsilon = 2$, $n_c = 15$, K-means clustering, and Closest to centroid sentence selection scheme)

| **ROUGE Measure** | **DUC 2002 Dataset** | | | **DUC 2007 Dataset** | | |
|---|---|---|---|---|---|---|
| | **Euclidean** | **Cosine Dissimilarity** | **Angular** | **Euclidean** | **Cosine Dissimilarity** | **Angular** |
| ROUGE-1 | 0.38990 | 0.33850 | 0.34723 | 0.38312 | 0.37144 | 0.36757 |
| ROUGE-2 | 0.13472 | 0.08001 | 0.09180 | 0.07023 | 0.06398 | 0.05829 |
| ROUGE-L | 0.37033 | 0.32105 | 0.33089 | 0.34826 | 0.34494 | 0.34273 |
| ROUGE-SU4 | 0.17816 | 0.12697 | 0.13713 | 0.13064 | 0.12325 | 0.12342 |

### *4.5. Different sentence selecting schemes*

In all the combinations of different clustering strategies, different index vector generation and different number of clusters, the closest to the centroid scheme outperformed the longest sentence scheme. The results of one particular combination, where the distinction is evident has been tabulated in Table 7.

Table 7: ROUGE recall scores of summaries generated for using different sentence selection schemes ($n_\varepsilon = 2$, $n_c = 15$, K-means clustering, Euclidean distance measure and Closest to centroid sentence selection scheme)

| **ROUGE Measure** | **DUC 2002 Dataset** | | **DUC 2007 Dataset** | |
|---|---|---|---|---|
| | **Closest to Centroid** | **Longest Sentence** | **Closest to Centroid** | **Longest Sentence** |
| ROUGE-1 | 0.38990 | 0.37268 | 0.38312 | 0.35443 |
| ROUGE-2 | 0.13472 | 0.12762 | 0.07023 | 0.06252 |
| ROUGE-L | 0.37033 | 0.35624 | 0.34826 | 0.31867 |
| ROUGE-SU4 | 0.17816 | 0.16975 | 0.13064 | 0.11676 |

### *4.6. Comparison with state-of-art systems*

The performance of our proposed summarization system has been compared with other state-of-art systems, viz. ClusterCMRW, ClusterHITS, RankClus, Spectral-Analysis, Local-Rank, Global-Rank, and the system proposed in [Cai & Li, 2013]. Prior to that we identified

the parameter combinations which produced best results for DUC 2002 and DUC 2007 datasets. For DUC 2002 dataset the parameter combination is as follows: $n_\varepsilon = 2$, $n_c = 5$, K-means clustering, Euclidean distance measure, and closest to centroid sentence selection scheme. For DUC 2007 dataset the parameter combination is as follows: $n_\varepsilon = 2$, $n_c = 15$, K-means clustering, Euclidean distance measure, and closest to centroid sentence selection scheme. The best recall scores for these two datasets are shown in Table 8. It is clear from this table that, our system performed much better for DUC 2002 dataset in comparison to DUC 2007 dataset.

Table 8: Best recall scores for DUC 2002 and DUC 2007 dataset.

| **ROUGE Measure** | **DUC 2002 Dataset** | **DUC 2007 Dataset** |
|---|---|---|
| ROUGE-1 | 0.39298 | 0.38312 |
| ROUGE-2 | 0.14259 | 0.07023 |
| ROUGE-L | 0.37334 | 0.34826 |
| ROUGE-SU4 | 0.18551 | 0.13064 |

To compare the performance of our system with other clustering-based system, for DUC 2002 dataset we have taken ROUGE-1 and ROUGE-2 recall scores of ClusterCMRW and ClusterHITS systems from [Wan & Yang, 2008]. The comparison of performance of these systems with our system is shown in Table 9. It is clear that our system (shown in bold letters) performs much better than other clustering-based system. The superiority of our system is much evident from the difference in ROUGE-2 recall scores.

Table 9: Comparison with other cluster-based summarization systems on DUC 2002 dataset

| **System** | **ROUGE-1** | **ROUGE-2** |
|---|---|---|
| **Our System** | **0.39298** | **0.14259** |
| ClusterCMRW (Kmeans) | 0.38221 | 0.08321 |
| ClusterCMRW (Agglomerative) | 0.38546 | 0.08652 |
| ClusterCMRW (Divisive) | 0.37999 | 0.08389 |
| ClusterHITS (Kmeans) | 0.37643 | 0.08135 |
| ClusterHITS (Agglomerative) | 0.37768 | 0.07791 |
| ClusterHITS (Divisive) | 0.37872 | 0.08133 |

For DUC 2007 dataset, we have taken ROUGE-1 and ROUGE-2 recall scores of clustering-based systems from [Cai & Li, 2013]. The comparison of performance of these

systems with our system (shown in bold letters) is shown in Table 10. However, the comparison is not too fair as the other clustering-based systems mentioned in this table are designed for query-focused multi-document summarization task, and the task for DUC 2007 dataset was also to create summary from multiple documents which answers a specific query. However, the summarization system proposed here is not designed to address specific queries, rather it is aimed at producing generic summaries. As a consequence, the performance of the proposed system appears to be marginally less as shown Table 10. It is clear from the ROUGE values that the proposed system is still highly competitive.

Table 10: Comparison with other cluster-based summarization systems on DUC 2007 dataset

| **System** | **ROUGE-1** | **ROUGE-2** |
|---|---|---|
| **Our System** | **0.38312** | **0.07023** |
| Cai & Li (2013) | 0.41662 | 0.12013 |
| RankClus | 0.41047 | 0.11579 |
| Spectral-Analysis | 0.40906 | 0.10341 |
| ClusterHITS | 0.39816 | 0.09104 |
| Local-Rank | 0.39489 | 0.08848 |
| Global-Rank | 0.38759 | 0.08012 |

## 5. Conclusion and Future Work

The present paper proposes and examines a combination of strategies for multi-document text summarization. The summarization strategies are based on different clustering and sentence selection schemes. Additionally, the paper also examines the individual effects of different cluster parameters by varying their values. The effect of index vector with different numbers of `+1's and `-1's on the summary generation has also been analysed. The proposed summarization scheme has been evaluated on the DUC 2002 and DUC 2007 datasets. It is observed that index vectors with one '+1' and one '-1' produced the best summaries for all combinations of other parameters. It may also be observed that summaries generated by K-means clustering strategy is qualitatively better than those generated using EMGM clustering strategy. Similarly, it can be observed that the closest to centroid sentence selection scheme outperformed the selection scheme based on sentence length. Moreover, it is noticed that Euclidean distance measure outperforms cosine dissimilarity and angular distance measures. However, no such conclusion could be derived regarding the number of clusters to be generated for summarization task. The reason may be attributed to the fact that the number

of themes varies from document set to document set, and consequently, the optimum number of clusters depends upon the number of themes present in a document set.

The evaluation results on DUC 2002 dataset showed that our summarization system with $n_\varepsilon = 2,\ n_c = 5,$ K-means clustering, Euclidean distance measure, and closest to centroid sentence selection scheme outperformed state-of-art ClusterCMRW and ClusterHITS summarization systems. For DUC 2007 as the existing cluster-based summarization schemes have been developed for query-focussed summaries, it is expected that their performance in comparison with generic summaries will be better when measured using recall based metrics. In spite of that the ROUGE scores obtained for the summaries produced by the proposed system proposed are pretty close to the other clustering-based summarization schemes. We find the performance of the proposed scheme of using Random Indexing and sentence clustering quite motivating. We intend to extend the proposed scheme for generating query-focussed summaries in future.

**References**


[1] Cai, X. and Li, W., A Spectral Analysis Approach To Document Summarization: Clustering and Ranking Sentences Simultaneously, *Information Sciences*, Vol. 181, No. 1, pp.3816-3827, September 2011.

[2] Cai, X. and Li, W., Ranking Through Clustering: An Integrated Approach to Multi-Document Summarization, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, N0. 7, pp. 1424-1432, July 2013.

[3] Cao, Z., Wei, F., Dong, L., Li, S. and Zhou, M., Ranking with Recursive Neural Networks and Its Application to Multi-document Summarization, In *29th AAAI Conference on Artificial Intelligence*, February 2015.

[4] Carbonell, J. and Goldstein, J., The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335-336, August 1998.

[5] Chatterjee, N. and Mohan, S., Extraction-Based Single-Document Summarization Using Random Indexing, In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Vol. 2, pp. 448-455, 2007.

[6] Chatterjee, N. and Sahoo, P. K., Effect of Near-orthogonality on Random Indexing Based Extractive Text Summarization, *International Journal of Innovation and Applied Studies*, Vol. 3, No. 3, pp. 701-713, July 2013.

[7] Chatterjee, N. and Sahoo, P. K., Random Indexing and Modified Random Indexing Based Approach for Extractive Text Summarization, *Computer Speech and Language*, Vol. 29, No. 1, pp. 32-44, January 2015.

[8] Gong, Y. and Liu, X., Generic Text Summarization using Relevance Measure and Latent Semantic Analysis, In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pp. 19-25, 2001.

[9] Gustavsson, P. and Jönsson, A., Text Summarization Using Random Indexing and PageRank, In *Proceedings of the 3rd Swedish Language Technology Conference (SLTC-2010)*, Linköping, Sweden, 2010.

[10] Hartigan, J.A. and Wong, M.A., Algorithm AS 136: A K-Means Clustering Algorithm, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 100-108, 1979.

[11] Higgins, D. and Burstein, J., 2007, Sentence Similarity Measures for Essay Coherence, In *7th International Workshop on Computational Semantics (IWCS)*, Tilburg, Netherlands, 2007.

[12] Hong, K. and Nenkova, A., Improving the Estimation of Word Importance for News Multi-Document Summarization, In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, pp. 712–721, April 2014.

[13] Klema V. C. and Laub, A. J., The Singular Value Decomposition: Its Computation and Some Applications, *IEEE Transactions on Automatic Control*, Vol. AC-25, No. 2, pp. 164-176, April 1980.

[14] Landauer, T., and Dumais, S., A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge, *Psychological Review*, Vol. 104, No. 2, pp. 211-240, 1997.

[15] Landauer, T., Foltz, P., and Laham, D., An Introduction to Latent Semantic Analysis, *Discourse Processes*, Vol. 25, pp. 259-284, 1998.

[16] Lee, D. D. and Seung, H. S., Algorithms for Non-Negative Matrix Factorization, In *Advances in Neural Information Processing Systems*, pp. 556-562, 2001.

[17] Lin, C.-Y., ROUGE: A Package for Automatic Evaluation of Summaries, In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Vol. 8, July 2004.

[18] Mani, I., *Automatic Summarization*, Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001.

[19] Porter, M. F., An Algorithm for Suffix Stripping, *Program*, Vol. 14, No. 3, pp. 130-137, 1980.

[20] Radev, D. R., Jing, H., Styś, M., and Tam, D., Centroid-Based Summarization of Multiple Documents, *Information Processing and Management*, Vol. 40, No. 6, pp. 919-938, 2004.

[21] Redner, R. A. and Walker, H. F., Mixture Densities, Maximum Likelihood and the EM Algorithm, *SIAM Review*, Vol. 26, No.2, pp.195-239, April 1984.

[22] Sahlgren, M., An Introduction to Random Indexing, In *Methods and Applications of Semantic Indexing Workshop, 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, Denmark, 2005.

[23] Sahoo, P. K., *Extractive Text Summarization Using Random Indexing*, Ph. D. Thesis, Indian Institute of Technology Delhi, India, October 2014.

[24] Sandhaus, E., *The New York Times Annotated Corpus*, Linguistic Data Consortium, Philadelphia, PA, 2008.

[25] Schütze, H., *Word Space*, In *Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems (NIPS'93)*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 895-902, 1993.

[26] Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H. and Wu, T., RankClus: Integrating Clustering With Ranking for Heterogeneous Information Network Analysis, In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 565-576, March 2009.

[27] Wan, X. and Yang, J., Multi-Document Summarization Using Cluster-Based Link Analysis, In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 299-306, July 2008.

[28] Wang, D., Li, T., Zhu, S. and Ding, C., Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization, In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 307-314, July 2008.