Simultaneous Clustering and Noise Detection for Theme-based Summarization

[1,2]Xiaoyan Cai, [2]Renxian Zhang, [2]Dehong Gao, [2]Wenjie Li

[1]College of Information Engineering, Northwest A&F University
xiaoyanc@mail.nwpu.edu.cn

[2]Department of Computing, The Hong Kong Polytechnic University
{csxcai,csrzhang,csdgao,cswjli}@comp.polyu.edu.hk

Abstract

Multi-document summarization aims to produce a concise summary that contains salient information from a set of source documents. Since documents often cover a number of topical themes with each theme represented by a cluster of highly related sentences, sentence clustering plays a pivotal role in theme-based summarization. Moreover, noting that real-world datasets always contain noises which inevitably degrade the clustering performance, we incorporate noise detection with spectral clustering to generate ordinary sentence clusters and one noise sentence cluster. We are also interested in making the theme-based summaries biased towards a user's query. The effectiveness of the proposed approaches is demonstrated by both the cluster quality analysis and the summarization evaluation conducted on the DUC generic and query-oriented summarization datasets.

1 Introduction

The exponential growth in the volume of documents available on the Internet brings the problem of finding out whether a single document can meet a user's complex information need. In order to solve this problem, multi-document summarization, which reduces the size of documents while preserves their important semantic content is highly demanded. Most of the summarization work done till date follow the sentence extraction framework, which ranks sentences according to various pre-specified criteria and selects the most salient sentences from the original documents to form summaries.

In addition to sentence salience, the other fundamental issues that must be concerned in summarization are information redundancy and information diversity (Radev et al., 2002). When the given documents are all supposed to be about the same topic, they are very likely to repeat some important information in different documents or in different places of the same document. Therefore, effectively recognizing the sentences with the same or very similar content is necessary for reducing redundancy and covering more diverse informative content in a summary. This is normally achieved by clustering highly related sentences into topical themes. Summaries can then be produced, e.g., by extracting the representative sentence(s) from each theme cluster. Thus, good sentence clusters are the guarantee of good summaries in theme-based summarization.

It is also important to stress that the noise sentences are clearly observed in the DUC datasets, i.e., the benchmark datasets for use by the summarization community (Wei et al., 2009). Take the DUC2005 d301i document set, which talks about 'International Organized Crime', as an example. The sentence like 'This well-educated, well-spoken, cosmopolitan businessman is laughing all the way.' absolutely goes too far off the point, and it is considered as a noise sentence in the context of our study. The existence of noises will inevitably degrade the clustering performance. Noise detection for summarization which has been ignored previously is emphasized in this work. Our strategy is to detect the noises by mapping the textual objects (either sentences or words) to a new representation space where the features are more discriminative. Then all the identified noises are thrown into a single cluster called noise cluster and the summaries are generated from the other regular clusters alone.

Topical themes and noises are the inherent characteristics of documents. Without doubt, effective recognition of them provides a good basis for theme-based summarization. However, summaries generated in such a way are not guaran-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 491–499,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

teed to cater to the user's information need and therefore may not always be in line with his/her expectations. For example, if a user asks to "identify and describe types of organized crime that crosses borders or involves more than one country", the cases of international drug trafficking and international smuggling are definitely more relevant than the origin and the means of organized crime or the government's precautions, even though all of them are extractable main themes in the documents. That is why query-oriented summarization which requires concise information corresponding to a specific query has drawn much attention in recent years. Its challenge to theme-based summarization is how to better make use of the query information to guide the necessary clustering and ranking processes. We explore three approaches to incorporate the query information in theme-based summarization, including query-driven cluster ranking, query-embedding similarity measure and semi-supervised clustering.

The main contributions of this paper are threefold. (1) Noisy detection is incorporated into clustering for theme-based generic summarization. (2) Three approaches are explored to incorporate the query information into query-oriented theme-based summarization. (3) Thorough experimental studies are conducted to verify the effectiveness and robustness of the proposed frameworks and approaches.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 explains the noise detection enhanced sentence clustering approach. Section 4 then addresses the other necessary issues in generic and query-oriented theme-based summarization. Section 5 presents experiments and evaluation results. Section 6 concludes the paper.

2 Related Work

Depending on the purpose and the target user, summarization can be either generic or query-oriented. While a generic summary reflects the author's point of view with respect to the most important information in the documents, a query-oriented summary presents the information in the documents that is most responsive to a given query.

Normally, sentence ranking is the issue of most concern in summarization (either generic or query-oriented). With the advancement of information technologies and the explosion of information on the Internet, clustering has become increasingly important in text mining and knowledge discovery. Recently it has been successfully applied in theme-based (a.k.a. clustering-based) summarization.

In terms of the roles of clustering in summarization, one could take the advantage of the clustering results to select the representative sentences in order to generate diverse summaries. The typical examples of such use are C-RR and C-LexRank proposed by Qazvinian and Radev (Qazvinian and Radev, 2008), which selected the important citation sentences from the sentence clusters generated by a hierarchical agglomeration algorithm. Alternatively, the clustering results could be used to improve or refine the sentence ranking results. Most of the clustering-based summarization approaches are of this nature. For example, Wan and Yang (Wan and Yang, 2008) presented a cluster-based conditional Markov random walk model and a cluster-based HITS model to incorporate the cluster-level information into the process of sentence ranking. Wang et al. (Wang et al., 2008a) also proposed a language model to cluster and summarize documents simultaneously using non-negative factorization. In addition, Wang et al. (Wang et al., 2008b) applied symmetric matrix factorization to generating sentence clusters. Each sentence's score is based on the linear combination of two elements. One is the average similarity score between a sentence and all the other sentences in the same cluster. The other is the similarity between the sentence and the given query. Notice that this is the only work we could find that explored clustering for query-oriented summarization.

Another important problem that we'd like to emphasize here is the existence of noisy data in any real-world dataset. To the best of our knowledge, no related work in summarization has attempted to solve this problem. In this paper, we try to address this issue by borrowing ideas from the noise detection research in the data mining literature. Existing noise detection approaches fall into two main types. One considered the data points whose distances to all cluster centers exceeded a certain threshold as noises after clustering (Dave, 1999). This type of approaches mainly focused on reducing the influence of noises on the regular clusters, but not exactly on identifying and removing noises. In this sense, the clusters output were still the noisy clusters. The other type managed to obtain one or more regular clusters and a single noise cluster that contained all noises simultaneously during clus-

tering (Li et al., 2007). Therefore, this type of approaches was able to provide noise-free clusters. The approach we are interested in this work is of the second type.

3 Spectral Clustering with Noise Detection

Compared to the traditional clustering algorithms such as K-means and agglomerative clustering, the new clustering algorithms that emerged over the last few years such as spectral clustering have demonstrated excellent performance on some challenging tasks (Ding and Zha, 2011). The spectral clustering has many fundamental advantages. For example, it is able to obtain global optimal solution and adapt to sample spaces with any shape (Ng, Jordan and Weiss, 2001; Bach and Jordan, 2004; Yu and Shi, 2003). The algorithm is also very simple to implement. It can be solved efficiently by standard linear algebra methods (Luxburg, 2007) and can be applied on a dataset of high dimensions in the feature space and data space (Dhillon et al., 2004). Taking into account these advantages, we choose to use spectral clustering in this study.

Without exception, spectral clustering is also sensitive to noises like all the other clustering algorithms. The main reason leading to its failure on the noisy dataset is that the block structure of the affinity matrix is destroyed by noises (Li et al., 2007). A possible solution is to reshape the noisy dataset so that the block structure of the new affinity matrix can be recovered. In this paper, we incorporate noise detection with spectral clustering by mapping the text data points from their original feature space into a new feature space such that a noise cluster formed by all the noisy data points can be separated from the other regular clusters. Basically, noise detection enhanced spectral clustering involves normalized graph Laplacian construction, data re-representation and spectral embedding.

3.1 Normalized Graph Laplacian Construction

Let *G*=(*S, A*) be an undirected weighted graph. $S = \{s_1, s_2, \cdots, s_n\}$ is the set of nodes corresponding to the text points represented as the *m*-dimensional feature vectors, *m* is the total number of the words and *n* is the total number of the sentences in a given document collection. $A = [a_{ij}]_{n \times n}$ is a symmetric matrix where a_{ij} is the weight of the edge connecting the two nodes s_i and s_j in G ($i, j = 1,2,\ldots,n$), and it is measured by the cosine similarity between the s_i and s_j vectors. The graph Laplacian L of G is defined as $L = I - A$, where I is the identity matrix, and the normalized graph Laplacian $\bar{L}$ of G is defined as

$$\bar{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2} \quad (1)$$

where $D = [d_{ij}]_{n \times n}$ is a diagonal matrix with $d_{ii} = \sum_j a_{ij}$. A is called the affinity matrix and $\bar{A} = D^{-1/2} A D^{-1/2}$ the normalized affinity matrix.

3.2 Data Re-Representation

In order to achieve relatively compact sentence clusters and meanwhile separate the noises from them, we map the sentence nodes $\{s_i, s_2, \cdots, s_n\}$ to $\{p_1, p_2, \ldots, p_n\}$ in a new feature space with dimension equal to *n*. It is expected that the block structure of the new affinity matrix built on this new graph can be recovered. Now let's consider the following regularization framework

$$\Omega(P) = \| P - I \|_F^2 + \alpha \cdot \mathrm{tr}(P^T K_r^{-1} P) \quad (2)$$

where $\| \cdot \|_F$ denotes the Frobenious norm of a matrix, $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, α is a positive regularization parameter controlling the trade-off between the two terms. K_r is a graph kernel (e.g., $K_r = \bar{L}^{-1}$). K_r^{-1} is the inverse of K_r if it is non-singular or is the pseudo-inverse of K_r if it is singular. $P = [p_1, p_2, \ldots, p_n]_{n \times n}^T$ is the new representation of the sentence set we would like to have after mapping. The optimal P can be obtained by minimizing $\Omega(P)$, i.e., $P^* = \arg\max_P \Omega(P)$. It is easy to see that the Equation (2) is strictly convex, so we could use the derivative of Equation (2) with respect to P to get the minimum of $\Omega(P)$, i.e.,

$$P^* = (I + \alpha K_r^{-1})^{-1} \quad (3)$$

Then the new representation of s_i is p_i^* (i.e. $P^*(i,\cdot)^T$, the *i*-th row vector of P^*).

3.3 Spectral Embedding

Given $P^* = \{p_1^*, p_2^*, \cdots, p_n^*\}$, i.e., the optimal representation of $S = \{s_i, s_2, \cdots, s_n\}$ in the new feature space, we can construct a new normalized

sentence graph Laplacian $\tilde{L}$. Let $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ be the eigenvalues of $\tilde{L}$ with the corresponding eigenvectors $v_1, v_2, \ldots, v_n$. Assume the cluster number k is known, let $V = [v_1, v_2, \ldots, v_k]_{n \times k}$. We normalize each row of V to unit length, resulting in a new matrix $\overline{V}$. The resultant row vectors correspond to the original sentence points and K-means clustering is performed on them. Then s_i is assigned to the cluster l $(1 \le l \le k)$ if and only if the i-th row vector in $\overline{V}$ (i.e., $\overline{V}(i,\cdot)$) is assigned to cluster l.

For each generated cluster, we compute the average distance between the sentence points in it and the origin. The cluster with the smallest average distance is taken as the noise cluster. The other clusters are considered as the regular clusters.

3.4 Cluster Number Estimation

Recall that spectral clustering requires a predefined cluster number k. To avoid exhaustive search for a proper cluster number for each document set, we employ the automatic cluster number estimation approach introduced in (Li et al., 2007) to predict the number of the expected clusters. Given the new normalized sentence graph Laplacian matrix $\tilde{L}$ and its eigenvalues $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$, the optimal number of clusters k^* is defined as

$$k^* = \arg\max_k \{\lambda_{k+1}(\tilde{L}) - \lambda_k(\tilde{L})\} \quad (4)$$

where $\lambda_i(\tilde{L})$ is the i-th smallest eigenvalue of $\tilde{L}$.

4 Theme-based Summarization

4.1 Generic Theme-based Summarization

Once the sentence clusters are obtained, the summary sentences are then extracted from the original documents according to the ranks of the ordinary clusters they belong to and their ranks within the assigned clusters.

The ranking score of each regular sentence cluster is formulated as:

$$\gamma(C_{S_i}) = \frac{Score(C_{S_i})}{\sum_{i=1}^{K} Score(C_{S_i})} \quad (5)$$

$$Score(C_{S_i}) = \frac{\sum_{j=1}^{m} C_{S_i}(j) \cdot S(j)}{\sqrt{\sum_{j=1}^{m} (C_{S_i}(j))^2} \cdot \sqrt{\sum_{j=1}^{m} (S(j))^2}} \quad (6)$$

where $Score(C_{S_i})$ indicates the cosine similarity between an regular sentence cluster C_{S_i} and the whole document set S for generic summarization. K is the total number of the ordinary sentence clusters identified; m is the number of the words in the whole document set. $\gamma(C_{S_i})$ is the normalized value of $Score(C_{S_i})$, where $\gamma(C_{S_i}) \in [0,1]$ and $\sum_{k=1}^{K} \gamma(C_{S_i}) = 1$.

Within each regular sentence cluster, any reasonable ranking algorithm, can be applied to rank the sentences. In view of the successful application of PageRank-like algorithms in sentence ranking, LexRank (Qazvinian and Radev, 2008) is adopted in our work. Considering sentence position also provides important information for generic summarization, we multiply a weight to the rank score of each sentence. The weight of a sentence in a document is 1/n, where n is the total number of the sentences in the document. This is a normal practice in generic summarization, which follows the hypothesis that the first sentence in a document is the most important and the importance decreases as the sentence gets further away from the beginning. The summaries are then generated by choosing the most salient sentence from the most salient regular cluster to the least salient regular cluster, then the second most salient sentences from the regular clusters in descending order of rank, and so on.

4.2 Query-Oriented Theme-based Summarization

For query-oriented summarization, the query's influence can be reflected in any process of ranking or clustering. Considering the focus of this study is the application of clustering in summarization, we explore three query-based approaches that center on the ranking process or the result of clustering.

4.2.1 Query-Driven Cluster Ranking

The process of it is similar to the generic theme-based summarization, except that the $Score(C_{S_i})$ is formulated as

$$Score(C_i) = \frac{\sum_{j=1}^{m} C_{S_i}(j) \cdot Q(j)}{\sqrt{\sum_{j=1}^{m} (C_{S_i}(j))^2} \cdot \sqrt{\sum_{j=1}^{m} (Q(j))^2}} \quad (7)$$

which indicates the cosine similarity between a sentence cluster and a given query *Q*. Moreover, the sentence position information can not be considered in query-oriented summarization.

As the query's influence is only reflected in the process of ranking in this approach, we argue that the query's influence can be not only reflected in the process of ranking, but also reflected in the process of clustering. We explore two query-based approaches that center on the result of clustering.

4.2.2 Query-Embedding Similarity Measure

Sentence clustering requires the affinity matrix that is built upon the cosine similarity between the two sentences. On top of query-driven cluster ranking, we further consider the query-driven sentence clustering by defining a new query-embedding similarity measure that biases inter-sentence relationships towards pairs of sentences possessing the same concepts expressed in the query.

The idea is intuitive. The *m*-dimensional sentence vector $s_i = (s_{i1}, s_{i2}, \ldots, s_{im})$ is mapped onto the *l*-dimensional query vector $Q = (q_1, q_2, \ldots, q_l)^T$ (l is the number of query terms). As the number of the words contained in query is much smaller than the number of the words contained in the document collection, to avoid problems resulting from exact word match, we propose to use the synsets in WordNet to map the sentence vector onto the query vector. Given a sentence s_i, for each q_t ($1 \le t \le l$), the weights of s_{ij} ($1 \le j \le m$) that are in the same synset as q_t are accumulated and contribute to be the weight of q_t. Consequently, the cosine similarity between the two sentences can be defined in the query vector space. We call it the query-embedding similarity.

4.2.3 Query-Supervised Clustering

Clustering is typically unsupervised. In the case that some limited prior knowledge is available, one can use the knowledge to "guide" the clustering process. This is called semi-supervised clustering. Inspired by this idea, we make use of the query information to supervise sentence clustering. It is expected that the sentences that correspond to certain aspects of the query will be grouped together forming the query-relevant clusters and the query-non-relevant sentences will be grouped together while the other noisy sentences fall into the noise cluster.

For this purpose, we adopt semi-supervised spectral clustering with pairwise constraints proposed by (Kamvar, Klein and Manning, 2003). We regard each query sentence as a seed for a query-relevant cluster and a sentence from the document collection which does not contain any word in the query sentences is selected to be a seed of the noise cluster. Then from the remaining sentences in the document collection, the one that has the highest cosine similarity to a seed is selected to construct a must-link constraint with that seed. Once a sentence is selected for a cluster, it cannot be assigned to the other clusters any more. Thus it can be naturally used to construct the cannot-link constraints with the other seeds.

As the query sentences are involved in clustering with this approach, the affinity matrix becomes $A = (a_{ij})_{(n+r)\times(n+r)}$, where r is the number of the sentences in the given query. Normally a_{ij} is defined as the cosine similarity between the two sentences s_i and s_j. Specially, $a_{ij} = 1$ is assigned to each pair of must-link constraint, indicating that the corresponding two sentences have to be in the same cluster. Similarly, $a_{ij} = 0$ is assigned to each pair of cannot-link constraint, indicating that the corresponding two sentences must not be in the same cluster. Then spectral clustering is applied based on this constrain-affinity matrix. We generate summaries from those clusters containing the query sentence(s). Other clusters are assumed to be either the query-non-relevant cluster(s) or the noise cluster.

4.3 Redundancy Control in Summary Generation

Since the number of documents to be summarized can be very large, information redundancy can be quite serious in the generated summaries. Redundancy control is necessary. We apply a simple yet effective way to choose summary sentences. Each time, we compare the current candidate sentence to the sentences already included in the summary. Only the sentence that is not too similar to any sentence already in the summary (i.e., the cosine similarity between them is lower than a threshold) is selected into the summary. The iteration is repeated until the length of sentences in the summary reaches the length limitation. In our experiment, the threshold is set to 0.9.

5 Experiments and Evaluation

We conduct a series of experiments on the DUC2004 generic summarization dataset and the DUC2007 query-based summarization dataset. According to task definitions, systems are required to produce a concise summary for each document set (without or with a given query description) and the length of summaries is limited to 665 bytes in DUC 2004 and 250 words in DUC2007.

A well-recognized automatic evaluation toolkit ROUGE (Lin and Hovy, 2003) is used for evaluation. We report two common ROUGE scores in this paper, namely ROUGE-1 and ROUGE-2, which base on the Uni-gram match and the Bi-gram match, respectively. Documents and queries are pre-processed by segmenting sentences and splitting words. Stop words are removed and the remaining words are stemmed using Porter stemmer.

5.1 Summarization Evaluation

To evaluate the performance of the noise detection enhanced spectral clustering approach, we compare the ROUGE scores of it with the ROUGE scores of the LexRank approach for generic summarization and query-oriented LexRank approach, which is a direct extension of LexRank in our clustering and ranking frameworks. That is, the sentence clusters are generated by the traditional spectral clustering algorithm first and then the sentences within each cluster are ranked with LexRank. With this approach, the cluster ranking and the summarization generation processes are exactly the same way as in our approaches. For LexRank and query-oriented LexRank approaches, we obtain the cluster number based on the normalized sentence graph Laplacian directly.

We choose $K_r = \overline{L}^{-1}$ and set α to 1000 for noise detection. Table 1 and Table 2 below illustrate the ROUGE results on the DUC2004 and DUC2007 datasets.

DUC2004	ROUGE-1	ROUGE-2
Noise Detection Enhanced	0.36325	0.07847
LexRank	0.36294	0.07351

Table 1. ROUGE Evaluation of Two approaches on DUC2004

DUC2007	ROUGE-1	ROUGE-2
Query-Driven Cluster Ranking	0.39351	0.09223
Query-Oriented LexRank	0.37589	0.07858

Table 2. ROUGE Evaluation of Two approaches on DUC2007

It is delighted to see that the noise detection enhanced clustering approaches consistently outperform the clustering approaches without noise detection in the both datasets. This demonstrates that removing noises can indeed benefit producing better sentence clusters that in turn can further enhance the performance of summarization.

5.2 Analysis of Cluster Quality

Our original intention to utilize noise detection enhanced spectral clustering is to hope to generate more accurate sentence clusters results by eliminating the negative impact of noises. In order to examine the quality of the generated sentence clusters, we define the following measure

$$quan = \sum_{k=1}^{K} \left(\frac{\min_{s_i \in C_k} sim(s_i, CC_{S_k})}{\sum_{l=1, l \neq k}^{K} \min_{s_i \in C_{S_l}, s_j \in C_{S_l}} sim(s_i, s_j)} \right) \quad (8)$$

where $\min_{s_i \in C_k} sim(s_i, CC_{S_k})$ denotes the distance between the cluster center and the border sentence in a cluster that is the farthest away from the center. The larger it is, the more compact the cluster is. $CC_{S_k} = \sum_{s_i \in C_{S_k}} s_i \Big/ |C_{S_k}|$ where $|C_{S_k}|$ is the size of C_{S_k}. $\min_{s_i \in C_{S_l}, s_j \in C_{S_l}} sim(s_i, s_j)$, on the other hand, denotes the distance between the most distant pair of sentences, one from each cluster. The smaller it is, the more separated the two clusters are. The distance is measured by cosine similarity. As a whole, the larger *quan* means the better cluster quality.

Table 3 and Table 4 below indeed clearly indicate the improved cluster qualities by removing noises and/or by making better use of the relationships among sentences and words. The ranges of the sentence clusters are also provided for reference.

DUC2004	*Quan*	*Cluster no.*
Noise Detection Enhanced	5.26	2-6
LexRank	4.73	2-7

Table 3. Cluster Quality Evaluation on DUC2004

DUC2007	*Quan*	*Cluster no.*
Query-Driven Cluster Ranking	4.79	3-6
Query-Oriented LexRank	4.18	3-7

Table 4. Cluster Quality Evaluation on DUC2007

5.3 Example of Effective on Noise Detection

Besides the quantitative evaluation, we also select the DUC2004 D30006 document set and DUC2007 D0702A document set to illustrate the advantages of enhancement with noise detection in generic and query-oriented summarization, respectively. The former document set contains 10 documents about 'Labor Dispute in National

Basketball Association', while the latter one contains 25 documents about 'Art and music in public schools' and the corresponding query is to 'Describe the state of teaching art and music in public schools around the word, indicate problems, progress and failures'.

Three relevant topical themes, including 'Employers' attitude', 'Employees' attitude' and 'Game Canceling' are mentioned in DUC2004 D30006 human summaries, 'Music and art education in the world', 'The problems of music and art education' and 'the progress and failure in the music and art education' are mentioned in DUC2007 D0702A, respectively.

For illustration, we compare the summaries generated by noise detection enhanced spectral clustering/query-driven cluster ranking and LexRank/query-oriented LexRank without noise detection. In order to provide better coherence of the generated summary, we group the sentences in the same cluster together in a paragraph and order them according to their ranking scores in that cluster.

If Feerick finds in favor of the owners, the reality of not being paid may spur the players to reach an agreement more quickly. In return for the concessions, the players want an increase in the minimum salary currently $ 272,500. **(Cluster 1: Employees' attitude)**
Larry Bird, in the Indiana countryside or inside Boston Garden, was a luminous exception to the governing rule. The proposal is similar to the luxury tax proposed by the union in 1995 during negotiations, but it would not be nearly as liberal. **(Cluster 2: Topic non-relevant)**
The National Basketball Association, embroiled in a labor dispute with its players, Tuesday canceled the first two weeks of the 1998 - 99season.**(Cluster 3: Game canceling)**

Table 5. System generated summary of DUC2004 D30006 using LexRank

But neither the players nor the owners are counting on the ruling by the arbitrator, John Feerick, to speed up negotiations, especially if Feerick finds in favor of the players, an award that could approach $800 million in salaries. **(Cluster 1: Employers' attitude)**
Next week, it will consider canceling the first-ever regular season games in league history. The NBA has already canceled the first two weeks of the regular season because of the labor dispute. **(Cluster 2: Game canceling)**
In return for the concessions, the players want an increase in the minimum salary currently $272,500 and creation of an average salary exception. More than 220 National Basketball Association players with guaranteed contracts will find out **(Cluster 3: Employees' attitude)**

Table 6. System generated summary of DUC2004 D30006 using noise enhanced spectral clustering

Yet many schools have overflowing classes, outdated textbooks, insufficient supplies and cuts in arts and sports. This is Inner City Arts, a nonprofit arts school that is both an enlightened model for arts education and a design landmark where education is embellished by architectural example. The Bingham Academy, in its third year, offers a five-week program for five disciplines: creative writing, dance, instrumental and vocal music, theater and visual arts. Given the national obsession with high-stakes tests, they reasoned, it made sense to promote art and music classes as a way to boost test scores. **(Cluster 1: The progress and failure in the music and art education)**
The design is also an object lesson in construction. Results of the study were released. In a variety of ways. Flamenco, for example, ties into social studies and language arts lessons on the history and culture of Spain. Test scores are rising. People want schools to teach conflict resolution by negotiation, not violence. Standardized tests have improved many American schools. And classical music in Cuba could, from this point of view, use a little rescuing. The last argument may be in trouble. **(Cluster 2: Topic non-relevant sentences)**
The requirements include four years of English; three years of math; two years of social science; two years of lab science, two years of foreign language; one year of visual or performing arts and one year of electives. Centralizing information about the arts is another matter. By some estimates, only 25 percent of American schools offer music programs as a basic part of the curriculum. Arts exchanges are only part of the business. **(Cluster 3: The problems of music and art education)**

Table 7. System generated summary of DUC2007 D0702A using query-oriented LexRank

Most had participated in Carnegie Hall's Linkup music education program, and it showed. The Roundabout Theater sends teaching artists to 40 classrooms in the city for 10 visits each. Artists from Lotus Music and Dance Studios in Chelsea work intensively with six schools across the city, including Public School 156, teaching students about the music and dance of different cultures. Delaine Easton, the State Superintendent of Public Instruction, has called for the restoration of arts education in California public schools. All arts curriculum was eliminated from the public schools. **(Cluster 1: The progress and failure in the music and art education)**
Given the national obsession with high-stakes tests, they reasoned, it made sense to promote art and music classes as a way to boost test scores. By some estimates, only 25 percent of American schools offer music programs as a basic part of the curriculum. The more prestigious University of California schools consider only the top one-eighth of the state's high school seniors, while California State University, dubbed the people's university, takes the top one-third. **(Cluster 2: The problems of music and art education)**
The rhythm, harmony and melodies of the music all create different perceptions and sensations within different regions of the brain. Areas of research will include new digital techniques for music, dance, storytelling and the visual arts. The theory that classical music makes the brain work better and they have some high-profile allies. To schedule this extra drill, students must drop an elective, like fine arts or gym. **(Cluster 3: The problems of music and art education, Benefit from the education)**

Table 8. System generated summary of DUC2007 D0702A using query-driven spectral cluster rankings

It is not difficult to conclude that the generated summaries using noise detection looks more informative than without using noise detection. We interpret the sentences of cluster 2 in Table 5 and the sentence of cluster 2 in Table 7 as the noise

sentences, considering they are not relevant to any main topical theme in the corresponding document set. Such kind of sentence is not observed in Table 6 and Table 8.

5.4 Comparison of Different Approaches to Integrating the Query Information

Different from generic summarization, the query information plays an important role in query-oriented summarization. In order to examine which way is more effective to integrate the query's influence into the clustering or ranking process, we further compare the query-based cluster-ranking and clustering approaches as introduced in Section 4.2. Meanwhile, we also implement the query-sensitive similarity measure introduced in (Tombros and Rijsbergen, 2001) for comparison, which calculates the similarity between s_i and s_j as

$$Sim(s_i, s_j) = \frac{\sum_{b=1}^{m} s_{ib} \cdot s_{jb}}{\sqrt{\sum_{b=1}^{m} s_{ib}^2 \cdot \sum_{b=1}^{m} s_{jb}^2}} \cdot \frac{\sum_{b=1}^{\max(l_m, l_Q)} c_b \cdot q_b}{\sqrt{\sum_{b=1}^{l_m} c_b^2 \cdot \sum_{b=1}^{l_Q} q_b^2}} \quad (9)$$

where l_Q is the query length of the query Q, l_m is the number of common words between s_i and s_j.

Table 9 below illustrates the ROUGE results on the DUC2007 dataset.

	ROUGE-1	ROUGE-2
Query-Driven Cluster Ranking	0.39351	0.09223
Query-Sensitive Similarity	0.39644	0.09537
Query-Embedding Similarity	0.39803	0.09698
Query-Supervised Clustering	0.40118	0.10125

Table 9. ROUGE Evaluation of Query-based Noise Detection Enhanced Approaches on DUC2007

We can observe from the above table that the query-embedding similarity approach and the query-supervised clustering approach clearly outperform the query-driven cluster ranking approaches. These results are expected. While the query-driven approach makes use of the query information in cluster ranking only, the other three approaches integrate the query information in both clustering and cluster ranking.

Beyond this, as a whole, the query-supervised clustering approach performs better than the query-embedding similarity approach. It can be interpreted if we look at the cluster generated. The query-supervised clustering approach is able to generate three types of clusters, i.e., query-relevant clusters, query-irrelevant clusters and a noise cluster and the summaries are generated merely from the query-relevant clusters. So the summaries generated are truly both query-relevant and theme-focused. In contrast, the query-embedding similarity approach can only differentiate the regular clusters from the noise cluster. Though the summaries are influenced by the query in some extent, the sentences in the generated regular clusters are not guaranteed to be relevant to the query.

It is also shown that the proposed query-embedding similarity measure has the advantage over the existing query-sensitive similarity measure, which simply multiplied the sentences-query similarity with the sentence-sentence similarity and thus the role of query is not as explicit as in query-embedding similarity.

To summarize, we believe that the high performance benefits from (1) Detecting and Removing Noise during Clustering i.e. removing noise sentences to enhance the clustering results and thus consequently improve the summarization performance; and (2) Guiding Sentence Clustering with Query for Query-oriented summarization, i.e., using the query information as the prior knowledge to supervise sentence clustering.

6 Conclusion

In conclusion, we propose noise detection enhanced spectral clustering to generate sentence clusters in this study. Moreover, we test the influence of query information for summarization generation. The experimental results on the DUC summarization datasets demonstrate the effectiveness and the robustness of the proposed approach. In particular the contribution of noise detection and the query information in the clustering process are clearly observed. In the future, we will add contextual information to sentences to further enhance the sentence clustering performance.

Acknowledgement

The work described in this paper was supported by an internal grant from the Hong Kong Polytechnic University (G-YH53) and UGC grant (PolyU 5230/08E).

References

Radev, D.R., Hovy, E., and Mckeown, K. 2002. *Introduction to the Special Issue on Summarization.* Computational Linguistics, 28: 399-408.

Wei F.R., Li W.J., Lu Q. and He Y.X. 2009. *Applying Two-Level Mutual Reinforcement Ranking in Query-Oriented Multi-Document Summarization*. Journal of the American Society for Information Science and Technology, 60(10):2119-2131.

Qazvinian V. and Radev D. R. 2008. *Scientific paper summarization using citation summary networks*. In Proceedings of 22nd COLING, pp.689-696.

Wan X. and Yang J. 2008. *Multi-Document Summarization Using Cluster-Based Link Analysis.* In Proceedings of 31st SIGIR, pp.299-306.

Wang D.D., Zhu S.H., Li T., Chi Y., Gong Y.H. 2008a. *Integrting Clustering and Multi-Document Summarization to Improve Document Understanding*. In Proceedings of 17th CIKM, pp.1435-1436.

Wang D.D., Li T., Zhu S.H., Ding Chris. 2008b. *Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization*. In Proceedings of 31st SIGIR, pp.307-314.

Dave RN. 1999. *Characterization and Detection of Noise in Clustering*. Pattern Recognist Lettter 12(11): pp.657-664.

Li Z.G., Liu J.Z., Chen S.F. and Tang X.O. 2007. *Noise Robust Spectral Clustering*. 11th ICCV, pp. 421-427.

Ding,C. and Zha H.Y. 2011. *Spectral Clustering, Ordering and Ranking*. Statistical Learning with Matrix Factorizations. 1st Edition.

Ng A.Y., Jordan M.I., Weiss Y. 2001. *On Spectral Clustering. Analysis and An Algorithm*. Advances in Neural Information Processing Systems 14: pp.849-856.

Bach F.R. and Jordan M.I. 2004. *Learning Spectral Clustering*. Advances in Neural Information Processing Systems 16, 1830-1847.

Yu S. X. and Shi J. 2003. *Multiclass Spectral Clustering*. 9th ICCV: pp.11-17.

Luxburg. U. 2007. *A Tutorial on Spectral Clustering*. Statistics and Computing. 17 (4): pp.395-416.

Dhillon I.S., Guan Y.Q. and Kulis B. 2004. *Kernel-K-means, Spectral Clustering and Normalzied Cuts*. 10th KDD, pp.551-556.

Lin, C. Y. and Hovy, E. 2003. *Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics*. HLT-NAACL2003, 71-78.

Tombros A. and Rijsbergen C.J. 2001. *Query-Sensitive Similarity Measures for the Calculation of Interdocument Relationships*. 10th CIKM, pp.17-24.

S.D.Kamvar, D.Klein, and C.D.Manning. 2003. *Spectral learning*. 18th IJCAI, pp.561-566.