Semi-Supervised Co-Clustering for Query-Oriented Theme-based Summarization

Libin Yang and Xiaoyan Cai
College of Information Engineering, Northwest A&F University, Xi n 712100, China

Abstract: Sentence clustering plays an important role in theme-based summarization which aims to discover the topical themes defined as the clusters of highly related sentences. However, due to the short length of sentences, the word-vector cosine similarity traditionally used for document clustering is no longer suitable. To alleviate this problem, we regard a word as an independent text object rather than a feature of the sentence and develop a noise detection enhanced co-clustering framework to cluster sentences and words simultaneously. We also explore a semi-supervised clustering approach to make the generated summary biased towards the given query. The evaluation conducted on the three DUC query-oriented summarization datasets demonstrates the effectiveness of the approaches.

Keywords: Co-clustering, semi-supervised, theme-based summarization

INTRODUCTION

Query-oriented summarization motivated by the Document Understanding Conferences (DUC) is of practical importance and increasing need due to the exponential growth of textual information on the WWW. Typically, it produces a concise and organized summary that can fulfill a user complex information need by extracting the most salient, representative and responsive sentences from the original documents.

There has been extensive research study on ranking sentences according to their salience for summarization (Gholamrezazadeh et al., 2009; Jin et al., 2010). Beyond that, there are other fundamental issues that must be considered, notably 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. Good sentence clusters are the guarantee of good summaries in theme-based summarization.

The key step in sentence clustering for theme-based summarization is to compute the similarity between two sentences. The traditional similarity measures based on word overlap alone that have been successfully used for document clustering could not be directly applied in sentence clustering, owing to the short length of sentences (Li et al., 2006). In order to provide a more reasonable sentence similarity value, recently several approaches have been proposed to convert the sparse word vector space into condensed concept vector space (Zhao et al., 2006; Islam and Inkpen, 2008; Cai and Li, 2010). These approaches in essence attempt to group the semantic-related words and then use word clusters to help generate sentence clusters. In other words, word clustering and sentence clustering are regarded as independent of each other. As a result, the sentence clustering performance is inevitably influenced by the word clustering result. To help alleviate this problem, we argue in this study that a word should be deemed as an independent text object instead of a feature of sentence. A co-clustering framework is developed to improve the performance of sentence clustering by allowing word to play an explicit role in sentence clustering.

It is also important to stress that the noise sentences (and noise words) are clearly observed in the DUC datasets, i.e., the benchmark datasets for use by the summarization community. Take the DUC2005 d301i document set, which talks about international Organized Crime as an example. The sentence like his 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 study. 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 guaranteed to cater to the user information need and therefore may not always be in line with his/her expectations. The challenge to theme-based query-oriented summarization is how to make better use of the query information to guide the necessary clustering and/or ranking processes. We explore a novel semi-supervised co-clustering approach for this purpose.

Co-clustering with noise detection: Compared to traditional K-means and agglomerative clustering, the newly emerged spectral clustering has many fundamental advantages, such as it is able to obtain global optimal solution and can be applied on a dataset of high dimensions in the feature space and data space etc. and in particularly it has demonstrated excellent performance on some challenging tasks (Ding, 2004). Taking these into account, 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. A possible solution is to reshape the noisy dataset so that the block structure of the new affinity matrix can be recovered. In this study, we borrow the ideas from the existing noise detection researches in the data mining literature. We incorporate noise detection with spectral clustering by mapping the text data points (sentence points or word points) from their original feature space into a new feature space such that a noise cluster formed by all the noise 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. Please refer to Reference (Li et al., 2007a) for the details.

The proposed co-clustering framework clusters sentences and words together in the same vector space. Formally, sentences, words and their relations are modeled as a graph \( G = (S, W, E) \), where \( S = \{s_1, s_2, \ldots, s_n\} \) and \( W = \{w_1, w_2, w_m\} \) are the sets of the nodes representing sentences and words respectively, \( n \) is the number of sentences and \( m \) is the number of words. Stop words are excluded. A word node is defined as the node that describes the WordNet (http://wordnet.princeton.edu) definition of the word in its most common sense. An edge in \( G \) connects a sentence to a word, such that \( E = \{\{s_i, w_j\}; s_i \in S, w_j \in W\} \). Each edge is associated with a weight signifying the cosine similarity between a sentence and a word definition. In such a way, the semantic relationship between words and sentences can be captured. The corresponding sentence-by-word matrix \( A_{SW} \) is constructed as:

\[
A_{SW} = [A_{SW}(i, j)]_{n \times m}
\]

where, \( A_{SW}(i, j) \) denotes the cosine similarity between the sentence \( i \) and the WordNet definition of the word \( j \). \( A_{SW} \) is interpreted as the affinity matrix of \( G \). The edges between words or between sentences are not considered in this framework.

The basic premise behind this framework is that sentences and words are assigned to \( K \) clusters simultaneously. That is, a sentence \( s_i \) should belong to the sentence cluster \( C_s(1 \leq i \leq K) \) if its association to the word cluster \( C_w \) is greater than its association to any other word cluster. Using the graph, the association of a sentence to a word cluster is the accumulated weight of the edges connecting it to all the words in that cluster:

\[
C_s = \left\{ s_i; \sum_{j \in C_w} A_{SW}(i, j) \geq \sum_{j \notin C_w} A_{SW}(i, j) \right\}
\]

(1 \( \leq s \leq K \) and \( f \neq l \))

So each sentence cluster is determined by the word clustering. Similarly given the sentence clusters \( C_s, \ldots, C_s_K \), the induced word clustering is given by:

\[
C_w = \left\{ w_j; \sum_{i \in C_s} A_{SW}(i, j) \geq \sum_{i \notin C_s} A_{SW}(i, j) \right\}
\]

(1 \( \leq s \leq K \) and \( f \neq l \))

This characterization is recursive in nature as the sentence clusters determine the word clusters, which in turn determines the better sentence clusters. Thus, the best word and sentence clustering would correspond to a partitioning of the graph such that the crossing edges between partitions have the minimum accumulated weight, which can be formulated as:

\[
cut(C_{w_1} \cup C_{w_2} \cup \ldots \cup C_{w_K}) = \min_{C_i, \ldots, C_k} \cut(C_i, \ldots, C_k)
\]

\( C_i \) is the joint cluster that consists of \( C_{w_i} \) and \( C_s(1, 2, \ldots, K) \)

We use the noise detection enhanced spectral clustering to solve the optimal problem and to generate \( K \)-1 regular clusters and one noise cluster. Notice that the regular clusters and the noise cluster contain both words and sentences. It is one of the advantages of this co-clustering framework that once the clustering is done, both the representative sentences and words of a cluster can be found and can be directly associated together.
To avoid exhaustive search for a proper cluster number, we employ the automatic cluster number estimation approach (Li et al., 2007b) to predict the number of the expected clusters, i.e.,

\[ K = \arg \max \{ \lambda_i \} \]

where, \( h = \max(m, n) \), \( \lambda_i \) and \( \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_h \) denote the normalized graph Laplacian matrix in the new feature space and its associated eigenvalues respectively and \( \hat{\lambda}_h(L) \) is the \( h \)-th smallest eigenvalue of \( L \).

**Semi-supervised co-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/word clustering. It is expected that the sentences/words that correspond to certain aspects of the query will be grouped together forming query-relevant clusters and the query-non-relevant clusters, while the other noise sentences/words will fall into the noise cluster.

For this purpose, we adopt semi-supervised spectral clustering with pair-wise constraints proposed by Reference (Kamvar et al., 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.

Taking sentence clustering as an example, 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, \( a_{ii} = 1 \) 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 constraint-affinity matrix. We generate summaries only 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. Semi-supervised word clustering operates in the same way.

<table>
<thead>
<tr>
<th>Table 1: ROUGE evaluation on DUC2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUC2005</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Semi-supervised co-clustering w/i noise detection</td>
</tr>
<tr>
<td>Co-clustering w/i noise detection</td>
</tr>
<tr>
<td>Co-clustering w/o noise detection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: ROUGE evaluation on DUC2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUC2006</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Semi-supervised co-clustering w/i noise detection</td>
</tr>
<tr>
<td>Co-clustering w/i noise detection</td>
</tr>
<tr>
<td>Co-clustering w/o noise detection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: ROUGE evaluation on DUC2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUC2007</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Semi-supervised co-clustering w/i noise detection</td>
</tr>
<tr>
<td>Co-clustering w/i noise detection</td>
</tr>
<tr>
<td>Co-clustering w/o noise detection</td>
</tr>
</tbody>
</table>

**Experiments on summarization:** The experiments are conducted on the three DUC query-oriented summarization datasets (DUC2005-DUC2007). The length of summaries is limited to 250 words in the 3 years. A well-recognized automatic evaluation toolkit ROUGE13 is used for evaluation. We report two common ROUGE scores here, i.e., ROUGE-1 and ROUGE-2.

Once regular sentence clusters are generated, the sentences can then be extracted from the original documents to form a summary according to the ranks of the regular clusters they belong to and the ranks of the sentences within their assigned clusters.

Let Q represent the query vector containing both the title and the narrative, the ranking score of each regular sentence cluster can be formulated as:

\[ r(C_k) = \frac{\text{score}(C_k)}{\sum_{i=1}^{K-1} \text{score}(C_i)} \]  

(4)

\[ \text{score}(C_k) = \frac{\sum_{j=1}^{m} C_k(j) Q(j)}{\sqrt{\sum_{j=1}^{m} C_k(i)^2} \sqrt{\sum_{j=1}^{m} Q(j)^2}} \]  

(5)

where, \( \text{score}(C_k) \) indicates the cosine similarity between the regular sentence cluster \( C_k \) and the given query Q. \( K-1 \) is the total number of the regular sentence clusters identified. \( r(C_k) \) is normalized \( \text{score}(C_k) \) such that:

\[ r(C_k) \in [0, 1] \]

\[ \sum_{i=1}^{K-1} r(C_i) = 1 \]

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, LexRank14 is adopted in
Table 4: System generated summary of DUC2005 d301i using semi-supervised co-clustering without noise detection

However, the Swiss will only co-operate with a foreign government if the crime being pursued is also a crime in Switzerland. Estimating the extent of organised crime is hampered by the lack of a broad definition in UK law of organised crime-such as that in the US law against racketeering. Law enforcement officers from nine African countries are meeting in Nairobi this week to create a regional task force to fight international crime syndicates dealing in ivory, rhino horn, diamonds, arms and drugs. **(Cluster 1: Countries involved in crime)**

Though his speech was a model of diplomatic balance, he stressed the many mistakes made by US law enforcement agencies in the original response to the growth of organised crime which permitted the expansion of a powerful, well-organised crime syndicate. He cited the recent example of Russian organised gangs working with the Italian Mafia to funnel a big drug consignment into the US. **CRIME WITHOUT FRONTIERS** By Claire Sterling Little Brown Pounds 18. Everyone has heard of the growth of crime in eastern Europe since the demise of communism. Quoting President Boris Yeltsin, Mr Freeh said that 'organised crime is trying to take the country by the throat'. **(Cluster 2: Countries involved in crime)**

"It's very scary," said one foreign drug expert, who requested anonymity. For instance, last September when Mr Libero Grassi, a Palermo businessman, was murdered for publicly refusing to pay Mafia extortion demands, the authorities amid much breath-beating agreed to introduce new anti-extortion measures. **(Cluster 3: Mafia extortion)**

our study. 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. Table 1 to 3 report the ROUGE results on the DUC2005, DUC2006 and DUC2007 datasets, where the three approaches are compared:

- Co-Clustering w/o Noise Detection
- Co-Clustering w/i Noise Detection (i.e., the query information is counted in cluster ranking as formulated in Eq. (4) but is not involved in clustering).
- Semi-Supervised Co-Clustering w/i Noise Detection (i.e., the query information plays the roles not only in cluster ranking, but also in sentence/word clustering).

Notice that ROUGE-2 is the primary DUC evaluation criterion. It is not surprising to find that the co-clustering with noise detection consistently significantly outperforms co-clustering without noise detection which demonstrates that removing noises can indeed benefit producing better sentence clusters that in turn further enhance the summarization performance. Most important, we are delighted to see that applying query-supervised clustering can further boost the performance. When comparing with other DUC participating systems, our approaches are ranked in the 3rd, 2nd and 2nd places among the 32 DUC2005, 35 DUC2006 and 32 DUC2007 participating systems, respectively.

Besides the quantitative evaluation, we also select the DUC2005 d301i document set to illustrate the advantages of enhancement with noise detection. This document set contains 41 documents about international Organized Crime. The corresponding query is to identify and describe types of organized crime, name the countries involved and identify the perpetrators involved with each type of crime. Three relevant topical themes are mentioned in human summaries, including types of crime, countries involved in crime and persons involved in crime. For illustration, we compare the summaries generated by noise detection enhanced co-clustering and co-clustering without noise detection, both incorporating semi-supervision into clustering. 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.

It is not difficult to conclude that the latter summary looks more informative than the former summary. Notice that we interpret the two sentences in the last paragraph in Table 4 as the noise sentences, considering they are not relevant to any main topical theme in the document set. While such kind of sentences is not observed in Table 5, the latter summary also effectively adds a necessary theme which has been missed in the former summary.

**CONCLUSION**

To summarize, we believe that the high performance achieved benefits from:

- Incorporating word inherent semantics in sentence clustering.
- Regarding words as independent text objects rather than features of sentences.
Detecting and removing noise during clustering.
Guiding sentence/word clustering with the given query.

ACKNOWLEDGMENT

This study was supported in part by the Fundamental Research Funds for the Central Universities under Grant No. Z109021108.

REFERENCES