Information Mining Based on Fusing Results of Multi-Perspective Cluster-based Summarizations

SARVNAZ KARIMI, ALI MOEINI, MAHMOUD R. HEJAZI

1Faculty of Engineering
Azad University, Science and Research Campus
Hesarak, Poonak, Tehran
Tel: +98 21 8497482-3
IRAN

2 Faculty of Engineering
University of Tehran,
Enghelab Ave., Tehran
Tel: +98 21 6406186
IRAN

3 Iran Telecom. Research Center, Info. Society Group
End of Kargar Shomali Ave., Tehran, P. O. Box: 14155-3961
IRAN

Abstract: In this paper, we propose a method to highlight the most important parts of documents, in order to help users to reach to their desired information in information space more beneficially. To do this, the results obtained from a multi-document cluster-based summarization system based upon different perspectives of summarization (such as query-based, general, etc.) are fused in a way that the final results are more informative with respect to the query, in comparison with the results obtained from each perspective of summarization individually.

Key-words: Information Mining, Information Retrieval, Fuzzy Document Clustering, Automatic Summarization

1 Introduction

Text information mining systems help users to find some related results in response to their needs expressed as queries, from large amount of data collections. The user expresses his/her need in a natural language description, and such systems will then mine through information space to find the related texts with respect to the query as beneficial as possible.

Some systems emerged for similar goals, use summarization, document clustering or both of them, namely, Scatter/Gather by Cutting et. al. in 1992 [1]. Scatter/Gather system, scatters the collection into a small number of document groups, or clusters, and presents short summaries of them to the user.

According to experiments of Prioli et al. [2] on Scatter/Gather system, they found that the subjects who used it were in much closer agreement about their understanding of topics and also provided richer descriptions of collection. However, more time to perform the search tasks was required with Scatter/Gather and the documents found were judged less relevant.
Hearst and Pederson [3] integrated Scatter/Gather with conventional search technology by clustering the results of a query. They found that relevant documents tended to fall mainly into one or two out of five clusters, and that the precision and recall were higher within best cluster than within the retrieval results as a whole. The implication of this finding is that all clusters of irrelevant documents can be filtered out once and the user can save significant time by looking only at the contents of seemingly promising clusters.

Russinov and Chen [4] proposed a system based on Hearst ans Pederson system by applying the relevance feedback mechanism on it. Their system, called Adaptive Search, assumes that top-ranked retrieved documents are relevant and uses them to augment the query with a relevance feedback ranking algorithm.

In the proposed system, we use from basic ideas of named systems. The system, first clusters query results and then summarize them; But, for extracting the results a new mechanism has been developed which is capable of extracting more useful and informative information in comparison to similar systems.

The main purpose of a summary is to give the reader an accurate and complete idea of the contents of the source and to present it to the user in a condensed form in natural language. In the new world of information, especially for the newswire or paper stories domain, producing a single-abstract is seldom enough to process a vast collection of information. This is why multi-document summarization is emerged whose goal is to identify what is common and what differs in a variety of the related documents and to remove redundant information from the summary [5].

Document clustering is one of the most important tasks in text mining. The textual, unstructured nature of documents makes it considerably more difficult in text mining in compare of its data mining counterparts [6]. Document clustering has been used to discover latent concepts in sets of unstructured text documents, and to summarize and label such collections. Clustering is inherently valuable in organizing and searching large text collections. Furthermore, clustering is useful for compactly summarizing, disambiguating and navigating the results retrieved by a search engine [7].

Several algorithms are proposed for document clustering. We applied a fuzzy clustering algorithm, for our purpose, called Sequential Fuzzy Competitive Clustering (SFCC). SFCC is an unsupervised heuristic algorithm which is computationally efficient and provides a good approximation of cluster prototype with cluster information on each dimension. It is an online clustering algorithm and is faster than K-means and most of the off-line clustering algorithms [8]. In [9], we have described the way it can be applied in document clustering.

We profit from clustering and summarization outcomes to discover the most informative parts of documents retrieved for a query. Document retrieval is done using SMART information retrieval system. The summarizer we have used is MEAD, which is a flexible summarization system under development at the university of Michigan.

The rest of the paper is organized as follows: In Section 2, we explain the summarization part of the system and how we reach to important parts of information. In section 3 system specification, implementation details, experimental results and evaluation measures are discussed. Section 4, concludes the paper.

2 Proposed Information Extraction Scheme Using Summarization

Automatic text summarization by a computer has been developed to alleviate the information overload problem. Text summarizations are based on statistical, linguistical and heuristic methods. In order to generate a summary, one has to identify the most important pieces of information from the document. The irrelevant information of a document is omitted and details of minimized relevant information are assembled into a compact coherent report.

Summarization process depends on features of a sentence. It means, different outputs are produced by different sentence scoring schemes that apply various features. Taking advantage of these features and their effect on extracting the information from texts, two summarization schemes can be defined, namely general summarization and query-based summarization:

• General document summarization which is based on centroids, highlights the most important sentences of documents of a cluster. In this step, the similarity of sentences with cluster centroids are considered. Three features called centroid, position of sentence in document and length of sentence are
considered in the method. In proposed method the sentence scores are computed using following formula:

General Score = Centroid + 0.5*Position  (1)

• Query-based summarization specifies those sentences that are strongly related to the user query. It considers zero weights for centroid and position features. This method considers the sentences similarity with the query. Sentence scores are computed using the following formula:

Query-based Score = SimTitle + SimDescription + SimNarrative  (2)

Where, SimTitle is the similarity of query title with sentence, SimDescription is the similarity of query description with sentence and SimNarrative is similarity of query narrative with sentence.

In the above schemes, length feature is essential for selection of summarized sentences. Sentences which are shorter than a predefined threshold are omitted. This threshold is chosen 10 words in our system. Our information mining system benefits from two above methods.

Although different summarization methods produce different results, there is an overlap between them. we conjectured document key-sentences that are in relation with query, should be in this overlap. The main idea of how we take advantage from summarization is explained in Fig. 1. The retrieved documents contain system source information, in sentence unit, and construct universal set U. Members of this set are given to the summarizer to create two other sets: G set, which contains sentences obtained from general summarization and Q set which contains sentences obtained from query-based method. The conjunction between sentences obtained from general summarization (G set in Fig. 1) and sentences obtained from query-based method (Q set in Fig. 1) are the desired abstract information need. In the case of empty set for G ∩ Q, top ranked sentences of Q can be considered for the system response. Top ranked sentences are those sentences which have higher scores among others. These scores are computed using formula (2).

It should be noted that this conjunction should be applied on top ranked sentences of each set. In other word, only high scored ones contribute in forming the sets G and Q. For this reason, a threshold for considering the sentences of these sets for applying in the proposed method is needed.

3 Implementation

The overall architecture of the system is shown in Fig. 2. User enters his/her query via user interface, and n top-ranked documents are retrieved for it using SMART information retrieval system. Top-ranked documents are those documents which are more relevant to the user query. The input of the system is a set of standard documents. SMART applies uniform preprocessing, stemming and weighting on the document collection. As a result, features related to each document are specified.

In SMART, users can specify different variants of weighting strategies based on term frequency and inverse document frequency. Among them a weighting scheme is chosen which may be computed using below formula:

\[
w_i = \frac{idf_i \times (0.5 + \frac{tfd_i}{2 \times \max tfd_j})}{\sqrt{\sum_{i=1}^{n} [idf_i \times (0.5 + \frac{tfd_i}{2 \times \max tfd_j})]^2}}
\]  (3)
Where, indexing weight $w_{ij}$, reflects the importance of each single-term $T_j$ in a document $doc_i$. $tf_{ij}$ is within-document term frequency, $df_j$ is collection-wide term frequency and $idf_j$ is computed as follows:

$$idf_j = \log \frac{N}{df_j} \quad (4)$$

Where, $N$ represents the number of documents.

It could be mentioned that this weighting scheme is called “ate” in SMART information retrieval system.

The next step is clustering retrieved documents using the SPCC algorithm. The clustered documents in rank order are given to the summarizer and cluster summaries are produced. In output formation step, the results are analysed to discover the most useful informations for the user, according to what explained in Section 2. User is either satisfied with the results or gives feedback to reformulate the query and submit the new one to the system.

In implementing the proposed system, MEAD summarization system has been used. This summarizer was first introduced by Radev et al. in 2000. MEAD is a centroid-based extractive summarizer that scores sentences based on sentence-level and inter-sentence features which indicate the quality of the sentence as a summary sentence. It then chooses the top-ranked sentences for inclusion in the output summary. In other words, the main idea behind MEAD is the use of centroid-based feature which identifies sentences that are highly relevant to an entire cluster of related documents [10],[11].

In order to score the sentences according to our desired formulas, we had to recompute the scores after getting the sentence features values from MEAD and perform desired process for choosing the sentences in output formation step.

### 3.2 Data Collection

The *Los Angeles Times* (LA) collection is used as a testbed for our experiments. This collection comes from the TREC (Text RETrieval Conference) initiative, which is very popular in Information Retrieval community. *LA* document collection consists of the full text of various newspaper and newswire articles plus government proceedings. It is a collection of randomly selected articles from 1989 to 1990. Each article is formatted with SGML-like tags, thus providing for an easy identification of titles and paragraphs. Beside this test collection, corresponding 150 topics are provided in TREC web site that include TREC6, TREC7 and TREC8 english test questions and relevance judgments. We used 60 of 150 queries to test our system performance. The above items are summarized in Table 1.

<table>
<thead>
<tr>
<th>Los Angeles Times Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Documents</td>
</tr>
<tr>
<td>Number of Queries</td>
</tr>
<tr>
<td>No. of Relevant Documents that Exist in Whole Collection</td>
</tr>
</tbody>
</table>

### 3.3 Evaluation Measures

In order to examine the system effectiveness, we have evaluated results using a combination of clustering quality and extracted information similarity to the user query which is computed using sentence scores acquired from query-based summarizations. The evaluation details and results of this evaluation is described below.

#### 3.3.1 Clustering Quality Measure

Ruger and Gauch have proposed an evaluation method in [12] to assess the quality of the clustering process based on human-expert data. A subset $R$ of retrieved documents from the collection C (Los Angeles Times collection) is partitioned to $n$ clusters called $R_1$, $R_2$…$R_n$ (see Fig. 3). $P_*$ is a parameter defined as the proportion of relevant documents retrieved to whole retrieved documents. The same proportion is computed for all clusters ($P_1$, $P_2$… $P_n$). The formulas for computing $P_*$ and $P_i$ are as follows:

$$P_* = \frac{n_r}{n_R} \quad \text{for } R \text{ subset} \quad (5)$$

$$P_i = \frac{n_{ri}}{n_{Ri}} \quad \text{for cluster } R_i \quad (6)$$

Where, $n_r$ is the number of relevant documents retrieved. $n_{ri}$ is number of relevant documents in cluster $R_i$, $n_R$ is number of retrieved documents for the query and $n_{Ri}$ is the number of documents in cluster $R_i$. 

Table 1. Collection statistics
The $p_i$ value varies from 0 to 1 as shown in Fig. 4 [12]. The values 0 and 1 are ideal for $p_i$. The worst case for clustering occurs when $p_i = P_*$. In this case the system can not help user to browse the retrieved documents.

![Fig. 2. System Architecture](image)

The quality measure $q_i$ for cluster $i$ is introduced such that $q_i$ interpolates linearly between the cases: ($q_i = 1$ for $p_i = 0$), ($q_i = 0$ for $p_i = P_*$) and ($q_i = 1$ for $p_i = 1$) (see Fig. 4). The $q_i$ is calculated as follows:

$$q_i = \begin{cases} 1 - \frac{p_i}{P_*} & \text{if } p_i < P_* \\ \frac{p_i - P_*}{1 - P_*} & \text{if } P_* \leq p_i \end{cases}$$

(7)

The weighted sum of these individual cluster qualities, weighted by the relative cluster size, give rise to an overall quality measure which is computed as follows:

$$q = \sum_i q_i \frac{|R_i|}{|R|} \quad \text{where} \quad q \in [0,1]$$

(8)

Where, $|R_i|$ is number of cluster $R_i$ (see Fig. 3) members and $|R|$ indicates the number of retrieved documents for the query.

![Fig. 3. Document Collection after Clustering](image)

![Fig. 4. Relation of parameters $p_i$ and $q_i$](image)
Averaging $q$ measure over all queries gives the behavior of the clustering algorithm in the context of a particular search engine (SMART in our case). Surely, higher values of $q$ show higher quality in clustering.

3.3.2 The Best Cluster Value Definition

Beside overall clustering quality for system evaluation, each cluster quality is needed separately. This quality is defined as the relevancy of the cluster to the query, which has direct relation with documents of that cluster. Since the $p_i$ parameter defined in formula (6) shows proportionality of relevant documents to whole documents of each cluster, it may be good candidate for expressing each cluster quality.

The higher value of $p_i$ shows better cluster or more useful one. The maximum value of $p_i$ between all clusters of each query shows “the best cluster”. The best cluster concept we used, is the cluster which has greatest $p_i$ value and is completely different from overall clustering quality introduced before.

3.3.3 System Evaluation

Evaluation of summarization system is inherently difficult due to multiple possible uses of a summary and difficulty of defining an ideal one [13]. That is why, there is no standard method for evaluation of summarization results. Most of the systems are evaluated by subject’s judgments. Even this method is in doubt. We have tried to simulate the user using cluster quality measure value as described below. In this way, we eliminate any bias that may be introduced from sources external to the system via user choices.

“The best cluster” obtained from clustering process as discussed before, contains more relevant documents in it and as a result, sentences extracted from these documents should be more helpful for the user. On the other hand, the similarity of one query to sentences of a summary can be computed using the scores of sentences according to different features including similarity to query title, description and narrative. Summation of sentence scores gives the overall summary score and this value is computed for every cluster of each query. The cluster with maximum score is known in this way and will be considered as “summary representative cluster”. If the summarization results are near to desired information, its chosen cluster should be the same as “the best cluster” or in a cluster with quality near to the best one. This is true if we trust on summarizer to give us the informative sentences of documents.

In this way, we can determine how much the summary representative cluster and the best one are near to each other. This distance is determined by quality difference of these two clusters. Less difference value shows better performance in our system.

3.4 Experimental Results

Using SMART 11/10 information retrieval system, 500 documents are retrieved for each query. The results of performing the algorithms on test collection, for 60 query are taken. The results have been evaluated according to the evaluation measure explained in previous sections. In the first step, quality of each cluster and overall clustering related to the query is computed. After summarization, “summary representative cluster” is known.

According to the results, conjunction of general and query-based summaries can be better than using absolutely query-based summarization results but there are some conditions that should be satisfied. We compared one-by-one results related to each query. The number of relevant information chosen for some queries has increased in proposed system but in some cases, sentences of cluster summary were not related or partially related to the query. In the case of increase, we observed that in 90 percent of queries, the overall clustering quality has been in an acceptable level. We can conclude that because the clustering quality has direct affect on cluster centroids, in the case of poor clustering, cluster centroid is deviated, so absolutely query-based summaries are more reliable than their combination with general summaries.

In order to have better basis for comparison of results, we considered a threshold for overall clustering quality. Between 60 queries, 23 of them produced acceptable results after clustering and their clustering quality was over 0.1. The summarized results obtained from our experiments are shown in Table 2.
### Table 2. System Evaluation Results

<table>
<thead>
<tr>
<th>Query</th>
<th>Clustering Quality</th>
<th>Best Cluster Quality</th>
<th>Cluster with Best Summary Quality</th>
<th>Difference of Cluster Quality of Best Cluster and Summary Representative Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Our system</td>
<td>Query-based</td>
<td>Our system</td>
<td>Query-based</td>
</tr>
<tr>
<td>1</td>
<td>0.242</td>
<td>0.033</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>2</td>
<td>0.112</td>
<td>0.195</td>
<td>0.195</td>
<td>0.105</td>
</tr>
<tr>
<td>3</td>
<td>0.106</td>
<td>0.138</td>
<td>0.090</td>
<td>0.081</td>
</tr>
<tr>
<td>4</td>
<td>0.182</td>
<td>0.033</td>
<td>0.033</td>
<td>0.016</td>
</tr>
<tr>
<td>5</td>
<td>0.115</td>
<td>0.496</td>
<td>0.339</td>
<td>0.339</td>
</tr>
<tr>
<td>6</td>
<td>0.152</td>
<td>0.156</td>
<td>0.156</td>
<td>0.156</td>
</tr>
<tr>
<td>7</td>
<td>0.146</td>
<td>0.157</td>
<td>0.157</td>
<td>0.082</td>
</tr>
<tr>
<td>8</td>
<td>0.398</td>
<td>0.089</td>
<td>0.089</td>
<td>0.008</td>
</tr>
<tr>
<td>9</td>
<td>0.157</td>
<td>0.025</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>10</td>
<td>0.229</td>
<td>0.058</td>
<td>0.033</td>
<td>0.008</td>
</tr>
<tr>
<td>11</td>
<td>0.254</td>
<td>0.113</td>
<td>0.040</td>
<td>0.024</td>
</tr>
<tr>
<td>12</td>
<td>0.150</td>
<td>0.260</td>
<td>0.260</td>
<td>0.260</td>
</tr>
<tr>
<td>13</td>
<td>0.149</td>
<td>0.189</td>
<td>0.139</td>
<td>0.139</td>
</tr>
<tr>
<td>14</td>
<td>0.102</td>
<td>0.228</td>
<td>0.213</td>
<td>0.131</td>
</tr>
<tr>
<td>15</td>
<td>0.137</td>
<td>0.114</td>
<td>0.114</td>
<td>0.056</td>
</tr>
<tr>
<td>16</td>
<td>0.116</td>
<td>0.123</td>
<td>0.123</td>
<td>0.098</td>
</tr>
<tr>
<td>17</td>
<td>0.118</td>
<td>0.254</td>
<td>0.138</td>
<td>0.138</td>
</tr>
<tr>
<td>18</td>
<td>0.141</td>
<td>0.113</td>
<td>0.113</td>
<td>0.106</td>
</tr>
<tr>
<td>19</td>
<td>0.129</td>
<td>0.180</td>
<td>0.165</td>
<td>0.115</td>
</tr>
<tr>
<td>20</td>
<td>0.109</td>
<td>0.187</td>
<td>0.171</td>
<td>0.187</td>
</tr>
<tr>
<td>21</td>
<td>0.129</td>
<td>0.195</td>
<td>0.185</td>
<td>0.185</td>
</tr>
<tr>
<td>22</td>
<td>0.184</td>
<td>0.115</td>
<td>0.107</td>
<td>0.115</td>
</tr>
<tr>
<td>23</td>
<td>0.149</td>
<td>0.115</td>
<td>0.115</td>
<td>0.115</td>
</tr>
</tbody>
</table>

Second column in Table 2 shows overal clustering quality. Cluster quality of the cluster chosen as the best one is in third column. Cluster quality of the produced summary for both methods (the common method of query-based and our system) is in forth column. Summation of differences between “the best cluster” and “summary representative cluster” has been specified in last row of the table. We can see our method overcomes the query-based one due to less difference of acquired results from ideal ones.

### 4. Conclusion And Future work

A method for extracting useful information from documents is proposed in this paper. This method is based on document clustering and summarization algorithms. A collection of TREC, called Los Angeles Times is chosen as dataset. Corresponding queries and relevance judgments are considered to evaluate the results. The results show that conjunction of sentence scoring between general summary sentences and query-based summary sentences are more reasonable than each of them individually.

In this study, we have trusted to the summarizer results that are not trustworthy in general because no context is considered, while context is essential for reaching to meaningful information. In future, we plan to expand the idea to provide a mechanism for query navigation in the same time and evaluate it using real subjects to measure its effect on the retrieval results.

Beside strongly query related informations in documents, some other important subjects can be found in retrieved documents that may be interesting for the user. Specially when user does not know the details of information he/she wants, they can be more useful. we are trying to extract the most important parts of them in the subset of general summaries minus those which are a member of query-based summaries. These information should be seperated by the document they belong to them. As a result, these information can be appeared as subject of new query suggestions in the system. In query suggestion part of the system, we still have not had enough experiments for evaluation of the results of query suggestion part. This work is under research and development for future.
References


