Topic Decomposition and Summarization

Wei Chen*, Can Wang, Chun Chen, Lijun Zhang, and Jiajun Bu

College of Computer Science, Zhejiang University
Hangzhou, 310027, China
{chenw,wcan,chenc,zljzju,bjj}@zju.edu.cn

Abstract. In this paper, we study topic decomposition and summarization for a temporal-sequenced text corpus of a specific topic. The task is to discover different topic aspects (i.e., sub-topics) and incidents related to each sub-topic of the text corpus, and generate summaries for them. We present a solution with the following steps: (1) deriving sub-topics by applying Non-negative Matrix Factorization (NMF) to terms-by-sentences matrix of the text corpus; (2) detecting incidents of each sub-topic and generating summaries for both sub-topic and its incidents by examining the constitution of its encoding vector generated by NMF; (3) ranking each sentences based on the encoding matrix and selecting top ranked sentences of each sub-topic as the text corpus' summary. Experimental results show that the proposed topic decomposition method can effectively detect various aspects of original documents. Besides, the topic summarization method achieves better results than some well-studied methods.

Keywords: Non-negative Matrix Factorization, Topic Decomposition, Topic Summarization, Singular Value Decomposition.

1 Introduction

Users nowadays are overwhelmed by the vast amount of information on the Web. Although they can find information for a specific topic easily using search engines, they still have difficulty in finding more detailed aspects of a topic before reading dozens of Web documents returned. For example, it is a non-trivial task to make a comprehensive survey of a topic such as "9/11 attacks". Related reports may cover various aspects (i.e., sub-topics) including "attackers and their motivation", "the rescue attempts", "9/11 investigations", etc. Each sub-topic may further contain a set of related incidents, e.g., "9/11 investigations" has a series of related incidents along the timeline, such as "the NSA intercepted communications that pointed to bin Laden on Sep.11, 2001", "FBI released photos of the 19 hijackers on Sep.27, 2001", etc. Thus, discovering sub-topics and related incidents for a specific topic in a text corpus and summarizing them will greatly facilitate user's navigation in the corpus space.

The above problems can be partially solved by topic decomposition and text summarization, which was first proposed systematically by Chen and Chen[1]. Their solution is called TSCAN (Topic Summarization and Content ANatomy). TSCAN equals to latent semantic analysis (LSA) based on the singular value decomposition (SVD)[2]. We

* This work is supported by China National Key Technology R&D Program (2008BAH26B00).

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 440–448, 2010.

also study this problem in this paper. However, our solution is based on Non-negative Matrix Factorization (NMF)[3]. NMF has been demonstrated advantages over SVD in latent semantic analysis, document clustering [4]. In our work, we model the documents of a specific topic as a terms-by-sentences matrix. NMF is used to factorize the matrix into a non-negative sub-topic matrix and a non-negative encoding matrix. Each row of the encoding matrix is examined to extract incidents and their summaries. Summary for each sub-topic is generated by composing its incidents' summaries. We rank sentences by analyzing the encoding matrix, and the top ranked sentences of each sub-topic are selected as the summary for the text corpus.

2 Related Work

For a given temporal documents of a specific topic, TSCAN has following steps: Firstly, the documents are decomposed into a set of blocks. Then, a $m \times n$ terms-by-blocks matrix $\mathbf{A}$ is constructed. $A_{i,j}$ is the weight of term i in block j, which is computed by TF-IDF weighting scheme. The block association matrix $\mathbf{B} = \mathbf{A}^T\mathbf{A}$, it is factorized as follow:

$$\mathbf{B} \approx \mathbf{T}_r\mathbf{D}_r\mathbf{T}_r^T \qquad \mathbf{T}_r \in \mathbb{R}^{n\times r}, \mathbf{D}_r \in \mathbb{R}^{r\times r}, \tag{1}$$

where $\mathbf{D}_r$ is a $r \times r$ diagonal matrix where the diagonal entries are the top r eigenvalues of $\mathbf{B}$. And $\mathbf{T}_r$ is a $n \times r$ matrix in which each of the r columns represents a sub-topic. By examining the constitution of each columns of $\mathbf{T}_r$, the significant incidents of each topic aspect are detected and their summaries are generated. Then, the summary of the topic documents is obtained by combining all detected incident's summary.

We assume the SVD of the terms-by-blocks matrix $\mathbf{A}$ as follow:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \qquad \mathbf{U} \in \mathbb{R}^{m\times m}, \mathbf{\Sigma} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}, \tag{2}$$

where both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{\Sigma}$ is a diagonal matrix. The diagonal entries of $\mathbf{\Sigma}$ are the singular values of the matrix $\mathbf{A}$. Each column of matrices $\mathbf{U}$ and $\mathbf{V}$ are called left-singular and right-singular vectors. Then,

$$\mathbf{B} = \mathbf{A}^T\mathbf{A} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}(\mathbf{\Sigma}^T\mathbf{\Sigma})\mathbf{V}^T. \tag{3}$$

The squares of singular values of the matrix $\mathbf{A}$ (i.e., $\mathbf{\Sigma}^T\mathbf{\Sigma}$) are equal to the eigenvalues of the matrix $\mathbf{A}^T\mathbf{A}$ (i.e., $\mathbf{B}$) [5]. In LSA, the r largest singular values with corresponding singular vectors from $\mathbf{U}$ and $\mathbf{V}$ are used to approximation the matrix $\mathbf{A}$[2], i.e.,

$$\mathbf{A} \approx \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T \qquad \mathbf{U}_r \in \mathbb{R}^{m\times r}, \mathbf{\Sigma}_r \in \mathbb{R}^{r\times r}, \mathbf{V}_r^T \in \mathbb{R}^{r\times n}, \tag{4}$$

Then, $\mathbf{B}$ can be approximated as follow:

$$\mathbf{B} = \mathbf{A}^T\mathbf{A} \approx \mathbf{V}_r(\mathbf{\Sigma}_r^T\mathbf{\Sigma}_r)\mathbf{V}_r^T. \tag{5}$$

Because $\mathbf{\Sigma}_r^T\mathbf{\Sigma}_r$ are the top r eigenvalues of the matrix $\mathbf{B}$, the matrix $\mathbf{V}_r$ is equal to the sub-topic matrix $\mathbf{T}_r$ derived by TSCAN. That is, the sub-topics derived by TSCAN corresponds to the right singular vectors with most significant singular values of $\mathbf{A}$. In this paper, we focus on extractive multi-document summarization which are widely studied[6,7,8,9]. It extracts top significant sentences calculated by a set of ranking methods from the documents set.

3 The Proposed Solution Based on NMF

Given a pre-specified topic t, it is represented as $D = \{d_1, d_2, \ldots, d_i, \ldots\}$, where d_i is a document at time point i. We call various topic aspects as sub-topics and define them as $ST = \{st_1, st_2, \ldots, st_k, \ldots, st_r\}$. The series of incidents corresponding to sub-topic st_k is defined as $STI_k = \{sti_{k,1}, sti_{k,2}, \ldots, sti_{k,i}, \ldots, sti_{k,l}\}$. Our task of topic decomposition is to find out sub-topics ST, the incidents STI_k related to sub-topic st_k. Besides, we generate summaries for the text corpus D, sub-topics ST and incidents STI. Each document in D is decomposed into a sequence of sentences using sentence separation software provided by DUC[10]. 425 Rijsbergen's stopwords are removed and stemming is performed. The documents in D is represented by a $m \times n$ terms-by-sentences matrix $\mathbf{M}$, where m is the number of terms and n is the number of sentences respectively. $M_{i,j}$ is computed by TF-IDF weighting scheme. The terms set is defined as $T = \{t_1, t_2, \ldots, t_i, \ldots, t_m\}$ while the sentences set is $S = \{s_1, s_2, \ldots, s_j, \ldots, s_n\}$.

3.1 Topic Decomposition Based on NMF

Non-negative Matrix Factorization (NMF) is a matrix factorization method that generates positive factorization of a given positive matrix[3]. It represents object as a non-negative linear combination of part information extracted from plenty of objects and is able to learn parts of semantic features from text. Given the matrix $\mathbf{M}$, NMF decomposes $\mathbf{M}$ into a non-negative matrix $\mathbf{B}$ and a non-negative matrix $\mathbf{E}$ so that

$$\mathbf{M} \approx \mathbf{BE} \qquad \mathbf{B} \in \mathbb{R}^{m \times r}, \mathbf{E} \in \mathbb{R}^{r \times n}. \tag{6}$$

We can find out $\mathbf{B}$ and $\mathbf{E}$ by minimizing the following cost function:

$$\arg\min_{\mathbf{B},\mathbf{E}} \|\mathbf{M} - \mathbf{BE}\|_F^2, \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The above constrained optimization problem can be solved by continuously updating $\mathbf{B}$ and $\mathbf{E}$ until the cost function converges under the predefined threshold or exceeds the number of repetitions[3,4]. The update rules are as follows ($1 \le i \le m$, $1 \le j \le n$ and $1 \le k \le r$):

$$B_{i,k} \leftarrow B_{i,k} \frac{(\mathbf{ME}^T)_{i,k}}{(\mathbf{BEE}^T)_{i,k}} \qquad E_{k,j} \leftarrow E_{k,j} \frac{(\mathbf{B}^T\mathbf{M})_{k,j}}{(\mathbf{B}^T\mathbf{BE})_{k,j}}. \tag{8}$$

The r columns of $\mathbf{B}$ embed the so called sub-topics and each column of $\mathbf{E}$ is the encoding. We refer $\mathbf{B}$ as the sub-topic matrix and $\mathbf{E}$ as the encoding matrix. Each sentence s_j can be represented by a linear combination of sub-topics. i.e.,

$$\mathbf{m}_j = \mathbf{Be}_j, \tag{9}$$

where $\mathbf{m}_j$ is j-th sentence (i.e., j-th column of $\mathbf{M}$) and $\mathbf{e}_j$ represents the j-th column of matrix $\mathbf{E}$. The entry $B_{i,k}$ indicates that the degree of term t_i belongs to sub-topic k, while $E_{k,j}$ represents that the degree of sentence s_j associates with sub-topic k.

Because the sentences set $S = \{s_1, s_2, \ldots, s_j, \ldots, s_n\}$ is indexed chronologically, the row k of encoding matrix $\mathbf{E}$ (i.e., $e_{k,j}$, $1 \le j \le n$, we refer it as sub-topic encoding

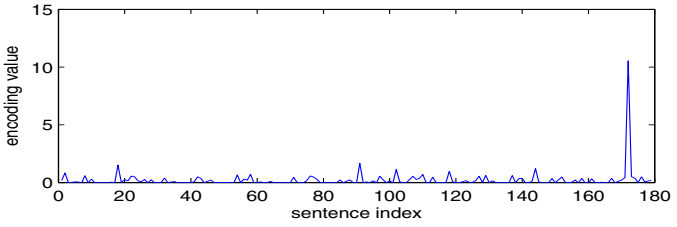

Fig. 1. A sub-topic's encoding vector of the document cluster 'd30001t' in DUC 2004

vector) also denotes the relative strength of sub-topic st_k along the timeline. Herein, a list of continuous elements of $e_{k,j}(1 \leq j \leq n)$ with high bursty values can be regarded as an incident related to sub-topic k. Fig. 1 shows a sub-topic encoding vector of document cluster 'd30001t' in DUC 2004[10] after applying NMF with $r = 10$. In Fig. 1, the encoding value is bursty around the sentence 170. It means that the sub-topic has a significant development around the sentence 170 (i.e., an incident breaks out).

The bursty detection problem is well studied in stream mining community [11]. Formally, given an aggregate function G(here is sum), a sliding window of size w and corresponding thresholds γ, the problem is to discover all these sub-sequences such that the function G applied to $e_{k,j:j+w-1}(1 \leq j \leq n - w + 1)$ exceeds threshold γ, i.e., check if

$$G(e_{k,j:j+w-1}) = \sum_{i=0}^{i=w-1} e_{k,j+i} \geq \gamma. \tag{10}$$

The threshold is set as $\gamma = mean(G(e_{k,j:j+w-1})) + \epsilon \times std(G(e_{k,j:j+w-1})), (1 \leq j \leq n-w+1)$, where $mean()$ and $std()$ are the average and standard deviation function respectively. We set ϵ as 3 and w as 7 in our experiments. Finally, the detected bursty sequences are recognized as incidents.

3.2 Topic, Sub-topics and Incidents Summarization

An interesting by-product of topic decomposition is that the produced information can also be used to generate summary. Lee et al. [9] also use NMF to do generic document summarization. In their work, the rank of each sentence is calculated as follows:

$$rank(s_j) = \sum_{k=1}^{r} (E_{k,j} \times weight(e_{k,1:n})), \tag{11}$$

$$weight(e_{k,1:n}) = \frac{\sum_{y=1}^{n} E_{k,y}}{\sum_{x=1}^{r} \sum_{y=1}^{n} E_{x,y}}. \tag{12}$$

The $weight(e_{k,1:n})$ is the relative relevance of k-th sub-topic among all sub-topics. Finally, the top-x sentences with highest rank are chosen as summaries. We refer this method as NMF in our experiments. As point out by [8], a good summary should contain as few redundant sentences as possible while contain every important aspects of the documents. However, the solution of Lee et al. [9] doesn't satisfy above requirements

sometimes. The top-x sentences with highest rank may belong to the same sub-topics and contain some overlapping information. In fact, most of the traditional summarization methods select sentences from different sub-topics [6,7]. We design a generic multi-document summarization method based on NMF (We refer it as INMF). Before going on, we give the definition of a sentence's main topic:

$$main\_topic(s_j) = \arg\max_k(E_{k,j}), \tag{13}$$

That is, the main topic of sentence s_j is the sub-topic with the maximum encoding value in column j of encoding matrix $\mathbf{E}$ [4]. The function $topic()$ returns the union of each sentence's main topic of a sentences set, e.g.,

$$topic(S) = main\_topic(s_1) \cup main\_topic(s_2) \cup \ldots \cup main\_topic(s_n). \tag{14}$$

The proposed multi-document summarization method INMF is described in Algorithm 1. Most of the summarization evaluations require the generated summaries in limited size or limited sentences number. In Algorithm 1, we limit the number of sentences of the summary. It can be easily revised to control the size of final summary. Different from [9], the INMF algorithm selects sentences with most significant ranking scores from different sub-topics in order to ensure the coverage and diversity of the summary. For each incident $sti_{k,i}$, we can straightforwardly choose the sentences with the largest or top-x encoding values of $sti_{k,i}$ as the summary. Then, the summary for sub-topic st_k can be generated by composing all the summaries of STI_k.

Algorithm 1. The proposed multi-document summarization method based on NMF

Input: $S = \{s_1, s_2, \ldots, s_n\}$; ns, the limitation of sentences number in summary
Output: a string array $SS = \{ss_1, ss_2, \ldots, ss_{ns}\}$, sentences set of summary
1: Sort sentences in S using equation 11: $rank(s_1) \geq rank(s_2) \geq \ldots \geq rank(s_n)$
2: $k = 1$; $TS = \emptyset$; // TS is the sub-topics set
3: for $k \leq ns$ do
4: $i = 1$;
5: for $i \leq size(S)$ do // size() returns the number of elements in set S
6: if $main\_topic(s_i) \notin TS$ then
7: $ss_k \leftarrow s_i$; $S = S - s_i$; $TS = TS \cup main\_top(s_i)$; $k++$; break;
8: end if
9: $i++$;
10: end for
11: if $size(TS) == r$ then // r is total sub-topics number
12: $TS = \emptyset$;
13: end if
14: end for

4 Experimental Studies

In the following, we first evaluate the proposed topic summarization method. And then, we give a case study of topic decomposition. The dataset of multi-document summarization task in DUC 2004[10] is used to evaluate the proposed methods.

4.1 Summarization Evaluations

We implement four baseline summarization systems: FORWARD(extracts the initial sentences of all documents of a topic); BACKWARD(generates summaries by selecting the end sentences of a topic); TSCAN; NMF(method of Lee et al., [9]). The number of sentences of summary generated by TSCAN is indeterminate. To ensure the comparison is fair, the evaluation procedure is as follows [1]: For each r, we firstly apply TSCAN to each document cluster to select a set of sentences as summary. Then, we use other methods to extract the same number of sentences for each r and document cluster. Both ROUGE-1 [12] and summary-to-document content coverage[1] metrics are used.

Table 1. Overall performance comparison of ROUGE-1 on DUC 2004

r	INMF	NMF	TSCAN	FORWARD	BACKWARD
2	0.31161	0.29707	0.23983	0.23875	0.18211
3	0.33475	0.32156	0.25342	0.25383	0.19766
4	0.36529	0.35522	0.27096	0.27092	0.22081
5	0.39042	0.38238	0.29288	0.29061	0.24342
6	0.39739	0.40410	0.30370	0.31152	0.25867
7	0.43594	0.42632	0.31636	0.32451	0.28100
8	0.46620	0.45518	0.33862	0.34409	0.29974
9	0.47680	0.47117	0.35014	0.35653	0.31159
10	0.48975	0.48382	0.36947	0.36348	0.32110

The overall performance comparison of ROUGE-1 on DUC 2004 is listed in Table 1. It shows that the two NMF based summarization methods get much better results than other methods for all r. This is because both the two NMF based methods try to cover all content as much as possible. However, TSCAN may not consider sub-topics successfully, FORWARD extracts beginning sentences and BACKWARD takes the end sentences. As r increase, TSCAN extracts more sentences as the summary. Because ROUGE is recall-oriented, the ROUGE-1 scores of all methods increase with the increasing of summary size as showed in Table 1. The proposed INMF method increase summary coverage by selecting sentences from different sub-topics explicitly. As a result, INMF outperforms NMF in most cases except $r = 6$.

A good summary should contain important aspects of original topic documents as much as possible [8]. Herein, we apply summary-to-document content similarity to measure the coverage of summary according to [1]. That is, given a document cluster of a specific topic and its summary which are represented by TF-IDF term vectors. It computes the average cosine similarity between each of the document clusters and its summary. The higher the similarity, the better the summary represents document cluster. We show the summary-to-documents similarity corresponding to table 1 in table 2.

In table 2, both INMF and NMF achieve much better results than TSCAN, FORWARD, BACKWARD for all r. It is easy to understand that all the three latter methods lose some information and the coverage is poor. However, Non-negative Matrix Factorization decomposes all of topic's information into r sub-topics, and the two NMF based summarization method extract the sentences with as much information as possible.

Table 2. Summary-to-document content similarity corresponding to table 1

r	INMF	NMF	TSCAN	FORWARD	BACKWARD
2	0.126703	0.123421	0.105427	0.103786	0.102524
3	0.128219	0.122003	0.107548	0.103958	0.104369
4	0.126229	0.121232	0.105550	0.104454	0.100763
5	0.127125	0.122199	0.108460	0.107260	0.102478
6	0.127987	0.122781	0.103569	0.104365	0.101695
7	0.128505	0.124848	0.102935	0.101614	0.102715
8	0.130945	0.126131	0.108045	0.105535	0.101850
9	0.127965	0.123313	0.106552	0.105038	0.099187
10	0.128130	0.124492	0.111100	0.109034	0.107870

Table 3. Sub-topic's description and the sentence id of each sub-topic's summary

ST id	sub-topic description	sentence id
1	Hun Sen and Ranariddh often clashed over power-sharing and the integration of guerrilla fighters from the crumbling Khmer Rouge.	74,119
2	King Norodom Sihanouk called Ranariddh and Sam Rainsy to return to Cambodia and wanted to preside over a summit meeting of the three party leaders.	20,46,56,57
3	Norodom Ranariddh and Sam Rainsy, citing Hun Sen's threats to arrest opposition figures, said they could not negotiate freely in Cambodia.	3,30,141,163
4	In July election, Hun Sen's party collected 64 of the 122 parliamentary seats, but was short of the two-thirds majority needed to set up a new government.	8,25,44,85,111
5	The violent crackdown in Cambodia, at least four demonstrators were killed.	40,64,69
6	Hun Sen and Ranariddh agreed to form a coalition that leave Hun Sen as sole prime minister and make Ranariddh president of the National Assembly.	83,123,152,175
7	People's Party criticized the resolution passed earlier this month by the U.S. House of Representatives.	59
8	King Norodom Sihanouk praised agreements by Cambodia's top two political parties previously bitter rivals to form a coalition government.	172
9	Sam Rainsy wanted to attend the first session of the new National Assembly on Nov. 25, but complained that his party members' safety.	160
10	The Cambodian People's Party gave a statement about supporting the police action of violent crackdown to protesters.	70

Besides, the proposed INMF summarization method explicitly tries to select sentences belong to different topic aspects. That's why INMF outperforms NMF in all cases.

4.2 Topic Decomposition

The documents set 'd30001t' of DUC 2004 is used as a case study for topic decomposition. It includes 10 documents and 179 sentences about "political crisis in Cambodia in October 1998". The detailed description about each sub-topic and sentence id of its summary is showed in table 3. We manually compared each sub-topics' summaries with reference summaries('D30001.M.100.T.A', 'D30001.M.100.T.B', 'D30001.M.100.T.C' and 'D30001.M.100.T.D' with size 658, 661, 647 and 656 bytes respectively) created

by DUC assessors. For 'D30001.M.100.T.C' and 'D30001.M.100.T.D', the information coverage is 100%. The text "the opposition tried to cut off his access to loans" (total 51 bytes) in 'D30001.M.100.T.A' and "Opposition parties ask the Asian Development Bank to stop loans to Hun Sen's government"(total 87 bytes) in 'D30001.M.100.T.B' are lost in the generated summaries. Then, the average information coverage of the generated sub-topics' summary to the reference summaries is $((658 - 51)/658 + (661 - 87)/661 + 100 + 100)/4 = 94.75\%$.

Some sub-topics contain a series of incidents while others contain only one. For example, sub-topic 6 is about the process of forming a coalition government which contains several incidents. Sub-topic 8 has only one incident, which is about King Norodom Sihanouk's praise about the agreements by Cambodia's top two political parties. We also compare our results with TSCAN for topic decomposition with $r = 10$. TSCAN detects total 23 incidents. 5 incidents are the same (some incidents duplicate more than 2 times), with only 15 sentences left as the incidents' summary. The 15 sentences are from 7 documents while the incidents' summaries of our method are from all 10 documents. Besides, our method covers more aspects than TSCAN, e.g. sub-topic 5 and sub-topic 10 are not included in the result of TSCAN.

5 Conclusion

In this paper, we study the problem of topic decomposition and summarization for a temporal-sequenced text corpus of a specific topic. We represent the text corpus as a terms-by-sentences matrix and derive sub-topics by factorize the matrix using Non-negative Matrix Factorization. By analyzing the encoding matrix, we can detect incidents of each sub-topic and generate summaries for both sub-topics and their related incidents. The summary for the text corpus is generated by firstly ranking each sentences based on the encoding matrix, and then selecting most significant sentences from each sub-topics. Experimental results show that our method can effectively find out different topic aspects of a documents set and generate promising results in summarization.

References

1. Chen, C.C., Chen, M.C.: TSCAN: A Novel Method for Topic Summarization and Content Anatomy. In: Proc. of the 31st ACM SIGIR conference, pp. 579–586. ACM, USA (2008)
2. Deerwester, S., Dumais, S.T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
3. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
4. Xu, W., Liu, X., Gong, Y.H.: Document Clustering Based on Non-negative Matrix Factorization. In: Proc. of the 26th ACM SIGIR conference, pp. 267–273. ACM, USA (2003)
5. Strang, G.: Introduction to Linear Algebra. Wellesley Cambridge Press, Wellesley (2003)
6. Gong, Y.H., Liu, X.: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In: Proc. of the 24th ACM SIGIR conference, pp. 19–25. ACM, USA (2001)

7. Zha, H.Y.: Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In: Proc. of 25th ACM SIGIR, pp. 113–120 (2002)
8. Wan, X.J., Yang, J.W., Xiao, J.G.: Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In: Proc. of IJCAI, pp. 2903–2908. ACM, USA (2007)
9. Lee, J.H., Park, S., Ahn, C.M., Kim, D.: Automatic generic document summarization based on non-negative matrix factorization. Info. Processing and Management 45, 20–34 (2009)
10. Document Understanding Conferences (2004), http://www-nlpir.nist.gov/projects/duc/index.html
11. Vlachos, M., Meek, C., Vagena, Z., Gunopulos, D.: Identifying Similarities, Periodicities and Bursts for Search Queries. In: Proc. of ACM SIGMOD, pp. 131–142. ACM, USA (2004)
12. Lin, C.Y.: ROUGE: a Package for Automatic Evaluation of Summaries. In: Proc. of the Workshop on Text Summarization Branches Out, Barcelona, Spain, pp. 74–81 (2004)