MSSF: A Multi-Document Summarization Framework based on Submodularity

Jingxuan Li, Lei Li, Tao Li
School of Computing and Information Sciences
Florida International University
Miami, FL
{jli003,lli003,taoli}@cs.fiu.edu

ABSTRACT

Multi-document summarization aims to distill the most representative information from a set of documents to generate a summary. Given a set of documents as input, most of existing multi-document summarization approaches utilize different sentence selection techniques to extract a set of sentences from the document set as the summary. The submodularity hidden in textual-unit similarity motivates us to incorporate this property into our solution to multi-document summarization tasks. In this poster, we propose a new principled and versatile framework for different multi-document summarization tasks using the submodular function [8].

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Performance

Keywords: Submodularity, Summarization, Framework

1. INTRODUCTION

Typical multi-document summarization tasks include generic and query-focused summarization. Recently, new summarization tasks such as update summarization [1] and comparative summarization [10] have also been proposed.

Lin et al. [5] proposed attacking generic summarization problem using submodularity along with a greedy algorithm. In this paper, we propose a new framework for ***M***ulti-document ***S***ummarization using the ***S***ubmodular ***F***unction (MSSF) which extends the submodularity to not only generic summarization, but also query-focused, update, and comparative summarization, and applies an improved greedy algorithm proposed by Minoux [6] (which is faster than the greedy algorithm in Lin et al. [5]) to generate summaries.

2. MSSF

In this section, first of all, the definition of submodularity is given since it is the core of MSSF. Then we explore how to map the summarization problem into a budgeted maximum coverage problem which is based on submodularity. Finally, the submodular functions for various summarization tasks performed by MSSF are presented.

2.1 Submodularity

Let E be a finite set and f be a real valued nondecreasing function defined on the subsets of E that satisfies

$$f(T \cup \{\varsigma\}) - f(T) \leqslant f(S \cup \{\varsigma\}) - f(S), \qquad (1)$$

where S and T are two subsets of E, $S \subseteq T$, and $\varsigma \in E \setminus T$, such a function f is called **submodular** function [8]. A key observation is that submodular functions are closed under nonnegative linear combinations [4].

Task	Submodular Function
Generic	$f(S) = \sum_{s_i \in D \setminus S} \sum_{s_j \in S} sim(s_i, s_j)$ $- \sum_{s_i, s_j \in S, s_i \neq s_j} sim(s_i, s_j)$
Query-focused	$f(S, q) = f_G + \sum_{s_i \in S} sim(q, s_i)$
Update	$f(q, S_1, S_2) = f_G + \sum_{s_i \in S_2} sim(q, s_i)$ $- \sum_{s_i \in S_2} \sum_{s_j \in S_1} sim(s_i, s_j)$
Comparative	$f(S) = f_G$ $- \sum_{s_i \in S} \sum_{s_j \in OtherGroups} sim(s_i, s_j)$

Table 1: A quick summary of the submodular functions for different summarization tasks.

Notation	Meaning
D	Document Set
S	Summary
S_1	Summary for D_1
S_2	Summary for D_2
w_t	Weight of term t
s_i, s_j	Textual unit
sim	Similarity
q	Given query
f_G	General information coverage

Table 2: Notations.

2.2 Problem Formulation

The summarization can be modeled as a budgeted maximum coverage problem. The budgeted maximum coverage problem is described as: given a set of elements E where each element is associated with an influence and a cost defined over a domain of these elements and a budget B, the goal is to find out a subset of E which has the largest possible influence while the total cost does not exceed B. This problem is NP-hard [3]. However, Khuller et al. [3] proposed a greedy algorithm which sequentially picks up the element that increases the **largest possible** influence within the cost limit and it guarantees the influence of the result subset is $(1 - 1/e)$-approximation. It is easy to observe that submodularity resides in each "pick up" step.

Based on the definition of submodular function and the budgeted maximum coverage problem, we can derive the utilization of submodularity for multi-document summarization tasks.

The basic step of multi-document summarization is to extract a candidate sentence. If the budget B is the number of terms in the summary, the cost of each candidate sentence is the number of terms within it. A high quality summary should maximize the information coverage of the given document set, while minimize the redundancy. One of the most popular methods for serving these two purposes is MMR [2]. Hence, a MMR-similar definition for the quality of the current generated summary is given by

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

	DUC05	DUC06	TAC08 A	TAC08 B
Type of task	Query	Query	Query	Update
# of topics	50	50	48	48
# of docs per topic	25-50	25	10	10
Summary length (# of words)	250	250	100	100

Table 3: The brief description of the data sets.

$$f(S) = \sum_{s_i \in E \setminus S} \sum_{s_j \in S} sim(s_i, s_j) - \sum_{s_i, s_j \in S, s_i \neq s_j} sim(s_i, s_j), \quad (2)$$

where E is the whole sentence set, and $sim(s_i, s_j)$ is the similarity between the textual units s_i and s_j (the typical textual unit is sentence). Note that the first component of Eq.(2) is for information coverage and the second component is for redundancy removal. Both components are submodular, thus $f(S)$ is also submodular, since the linear combination of submodular functions is closed. The goal of multi-document summarization is to generate a summary which provides the largest possible quality within the budget. Hence, multi-document summarization problem can be modeled as a budgeted maximum coverage problem.

Eq.(2) has presented the submodular function for generic summarization task, thus we can define more submodular functions for different summarization tasks. All the submodular functions are shown in Table 1, and descriptions of the notations are given in Table 2. Notice that f_G can be presented by the functions for generic summarization. As we have obtained corresponding submodular function for each summarization task, the algorithm by Minoux [6] is utilized for extracting summaries which have the quality close to optimal.

3. EXPERIMENTS

Lin et al. [5] have already showed the feasibility of the generic summarization using submodularity. Thus, we conducted experiments on three new summarization tasks to evaluate our proposed MSSF based on submodular function.

3.1 Data sets

Table 3 shows the characteristics of all the datasets used for our experiments. All the tasks, except the comparative summarization, are evaluated by ROUGE, an widely used evaluation toolkit for document summarization.

	DUC05		DUC06	
	ROUGE-2	ROUGE-SU4	ROUGE-2	ROUGE-SU4
Average-Human	0.10236	0.16221	0.11249	0.1706
DUC Average	0.06024	0.11488	0.07543	0.13206
Random	0.04143	0.09066	0.04892	0.10083
LSA	0.04079	0.09352	0.05022	0.10226
SNMF	0.06043	0.12298	0.08549	0.13981
Qs-MRF	0.0779	0.1366	0.08917	0.14329
Wiki	0.07074	0.13002	0.08091	0.14022
MSSF	0.0731	0.12718	0.09193	0.14611

Table 4: Results on query-focused summarization.

3.2 Results

For the **query-focused summarization** task, we compared our method with some widely used and recently published methods: SNMF [9], Qs-MRF [11], and Wiki [7]. The empirical results are reported in Table 4. Our method achieves the best result.

For the **update summarization** task, Table 5 shows the comparative experimental results, "TAC Best" and "TAC Median" represent the best and median results from the participants of the TAC08 summarization track in the two tasks respectively according to the TAC08 report [1]. The experimental results demonstrate that MSSF leads to the competitive performance for update summarization.

	TAC08 A		TAC08 B	
	ROUGE-2	ROUGE-SU	ROUGE-2	ROUGE-SU
TAC Best	0.1114	0.14298	0.10108	0.13669
TAC Median	0.08123	0.11975	0.06927	0.11046
MSSF	0.08327	0.12109	0.09451	0.13180

Table 5: Results on update summarization.

For **comparative summarization**, we use the top three largest clusters of documents from TDT2 corpora to generate the summary by MSSF. From the summaries presented in Table 6, MSSF can extract discriminative sentences for all the topics.

Topic	MSSF
Iraq Issues	The arrival of U.S. Secretary of State Madeleine Albright could be an early test of **the accord iraq signed ten days ago with U.N. secretary-general Kofi Annan**.
Asia's economic crisis	Prueher addressed the army seminar in Manila and told delegates that **Asia's financial troubles** have affected **the United States' joint military activities with its Asian allies**.
Lewinsky scandal	In Washington, Ken Starr's grand jury continued its investigation of the **Monica Lewinsky matter**.

Table 6: A case study on comparative document summarization. The bold font is used to annotate the phrases that are highly related with the topics.

We also conduct experiments to evaluate the efficiency of the improved greedy algorithm used in MSSF. Compared with using general greedy algorithm, by applying the improved greedy algorithm, the running time of different summarization tasks is reduced by 10,000-100,000 milliseconds.

4. CONCLUSION

In this poster, we propose a principled and versatile framework for different multi-document summarization tasks including generic, query-focused, updated, and comparative summarization using the submodular function. In addition, it adopts an improved greedy algorithm which reduces the running time of generating the summary with the competitive summarization quality.

Acknowledgements

The work is partially supported by US National Science Foundation under grants IIS-0546280 and CCF-0830659.

5. REFERENCES

[1] H.T. Dang and K. Owczarzak. Overview of the TAC 2008 Update Summarization Task. In *Proc. of TAC*, 2008.

[2] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop on Automatic summarization*, 2000.

[3] S. Khuller, A. Moss, and J.S. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 1999.

[4] A. Krause and C. Guestrin. Near-optimal observation selection using submodular functions. In *Proc. of AAAI*, 2007.

[5] H. Lin and J. Bilmes. Multi-document Summarization via Budgeted Maximization of Submodular Functions. In *NAACL/HLT*, 2010.

[6] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. *Optimization Techniques*, 1978.

[7] V. Nastase. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proc. of EMNLP*, 2008.

[8] GL Nemhauser, LA Wolsey, and ML Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 1978.

[9] D. Wang, T. Li, S. Zhu, and C. Ding. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proc. of SIGIR*, 2008.

[10] D. Wang, S. Zhu, T. Li, and Y. Gong. Comparative document summarization via discriminative sentence selection. In *Proc. of CIKM*, 2009.

[11] F. Wei, W. Li, Q. Lu, and Y. He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proc. of SIGIR*, 2008.