

Modifier Graph Based Subtopic Mining

Haitao Yu

Fuji Ren

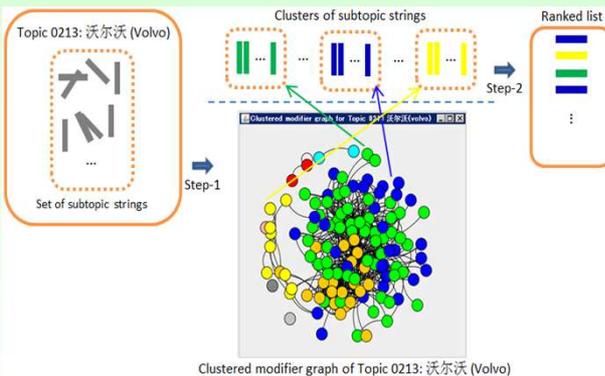
yu-haitao@iss.tokushima-u.ac.jp ren@is.tokushima-u.ac.jp

Faculty of Engineering, University of Tokushima, 2-1 Minamijyousanjima-cho, Tokushima, Japan

❖ Architecture

➤ Role-explicit Topic

Given a role-explicit topic, we firstly determine its kernel-object, then we construct the corresponding modifier graph (Step-1). We perform graph clustering on the induced modifier graph using the algorithm by Noack [4]. Each modifier cluster is viewed as a representation of a particular subtopic. Corresponding to each modifier cluster, we generate the clusters of subtopic strings based on the "child-parent" relationship between modifiers and subtopic strings. For example, the middle part of the Figure 2 shows the clustered modifier graph of an official topic. The modifier clusters are differentiated by colors. The clusters of subtopic strings have the same color with their child modifiers. Based on the clusters of subtopic strings, we generate the target ranked list.



➤ Role-implicit Topic

Because role-implicit topics generally express single information needs, we directly use the extracted subtopic strings. For the target ranked list, we sort the subtopic strings by their edit distance again the given topic and select the top-n (n is the required size of list L) to form the target ranked list.

Table 3. Role-implicit topics

ID	Topic
0245	央视主持人周涛简历 (resume of CCTV host ZhouTao)
0249	什么是京都议定书 (what is Kyoto Protocol)
0256	什么是RTF (what is RTF)
0270	陕西临潼农民发现秦始皇兵马俑 (the farmer in Lintong of ShanXi discovered the Terracotta Army)

❖ Results of TUTA1

Table 4. Result of Chinese subtopic Mining

Run	I-rec@10	D-nDCG@10	D#-nDCG@10
TUTA1-S-C-1A	0.4184	0.4686	0.4435
THUIS-S-C-1A	0.3881	0.4923	0.4402
THUIR-S-C-3A	0.3786	0.4987	0.4386
TUTA1-S-C-2A	0.4030	0.4655	0.4343
THUIS-S-C-4A	0.4036	0.4620	0.4328

Table 5. Result of English subtopic mining

Run Name	I-rec@10	D-nDCG@10	D#-nDCG@10
THUIR-S-E-1A	0.4107	0.3498	0.3803
THUIR-S-E-3A	0.3971	0.3492	0.3732
hultech-S-E-1A	0.3099	0.3991	0.3545
TUTA1-S-E-1A	0.1892	0.1756	0.1824
LIA-S-E-4A	0.1655	0.1740	0.1698
TUTA1-S-E-2A	0.1724	0.1569	0.1646

For the runs of Chinese subtopic mining, our proposed framework achieves the best performance in terms of I-rec and D#-nDCG. As most of the Chinese topics are role-explicit (92 out of 98 official topics), our modifier-graph based framework can incorporate a series of word-level knowledge, e.g., click information and semantic knowledge derived from the lexical ontology HowNet. Moreover, the proposed framework is robust to the sparseness problem commonly occurs when utilizing click information at a whole query level. However, our runs for English subtopic mining show a poor performance. We currently say that it is because of nonuse of a corresponding query log. We will further study this problem in our future work.

Acknowledgement

Our gratitude goes to the 10th NTCIR organizing committee for providing the platform of research-communication, and to their excellent work.

References

- [1] Yu, H.T., Ren, F.J.: Role-explicit Query Identification and Intent Role Annotation. In: Proceedings of the 21st CIKM, pp. 1163-1172 (2012)
- [2] Wen, J.R., Nie, J.Y., Zhang, H.J.: Clustering User Queries of a Search Engine. In: Proceedings of 10th WWW, pp. 162-168 (2001)
- [3] Gao, J.F., Yuan, W., Li, X., Deng, K.F., Nie, J.Y. Smoothing Clickthrough Data for Web Search Ranking. In: Proceedings of the 32nd SIGIR, pp. 355-362 (2009)
- [4] Noack, A.: Energy Models for Graph Clustering. Journal of Graph Algorithms and Applications, 11(2), pp. 453-480 (2007)