Generating High Level Descriptions from Cluster Transitions

Sophoin Khy[1] Yoshiharu Ishikawa[2] Hiroyuki Kitagawa[1,3]
[1] Graduate School of Systems and Information Engineering, University of Tsukuba
[2] Information Technology Center, Nagoya University
[3] Center for Computational Sciences, University of Tsukuba

1 Introduction

Clustering has been studied as a tool to group relevant information [2, 1, 6] and incremental clustering has also attracted great attention in recent years [4]. Clustering results are meaningful by themselves when they are browsed separately. As clustering is periodically performed and accumulated over time, clustering results however become inundated. It is difficult for users to interpret the overall information and trend from the large collection of clusters. Therefore it would be useful if we can provide the overall summary in a comprehensible and automatic way to users without them having to browse the abundant clustering results.

In this work, we explore a framework for generating high level descriptions from accumulated document clustering results. Given a user's query, our objective is to automatically present high level descriptive sentences as a summary from the entire document clustering results based on cluster transitions. It accommodates a search mechanism which allows users to specify what they want to know in the form of a query. Descriptive sentences are sentences which is a summary of the overall trend in the clustering results. They depend on the query results . Such sentences are expected to help user gaining insight of when and how a topic occurs in the underlying clustering results. A user is, for example, interested in knowing the development on the topic related to "Asashoryu". Descriptive sentences involving "Asashoryu" would be, for example, "'Asashoryu' appeared on Oct 1 to Oct 5, 2007, reappeared on Jan 6, 2008 and became very hot on Oct 3, 2007."

2 Related work

Spiliopoulou et al. proposed an approach called MONIC to model and track cluster transitions on clustering results at consecutive time points [5]. A cluster transition at a given time point is a change experienced by a cluster that has been discovered at an earlier time point. MONIC detects both internal and external transitions of clusters. Our approach is based on the external transitions of MONIC.

T-Scroll [3] is an information visualization interface tool to visualize the overall trend of a time series of documents. It provides scrollable interface, displays links between clusters and allows users to explore more detailed information such as contents of documents in a cluster. The idea and target problem in T-Scroll is related with our approach. While T-Scroll provides intuitive tool enabling users to see topics and how they change over time, our approach aims at providing informative summary of the overall trend of the underlying clustering results corresponding to the query by using search mechanism and descriptive sentences.

3 Proposed approach

The target data in this work is a collection of accumulated document clustering results $D_1, \ldots, D_n$ at consecutive time points $T_1, \ldots, T_n$. Let $S_1, \ldots, S_n$ be document sets in which S_i is the document set from which the document clustering result D_i were generated. S_i is assumed to be overlapped with S_{i+1}, i.e., $S_i \cap S_{i+1} \neq \emptyset$. However, all clusters $C_{i1}, \ldots, C_{in}$ for $D_i (1 \leq i \leq n)$ are assumed non-overlapped, i.e., $\forall p \neq q, C_p \in D_i, C_q \in D_i, C_p \cap C_q = \emptyset$. In order to generate high level descriptive sentences given a query, our approach proposes the following procedures:

1. Detect cluster transitions from the whole collection of document clustering results (Section 3.1).
2. Given a query, search for matched clusters to the query using cluster transitions (Section 3.2).
3. Generate descriptive sentences from search results (Section 3.3).

3.1 Detection of cluster transitions

In this work, we extend the cluster transition approach proposed in MONIC [5]. A cluster transition at a given time point is a change experienced by a cluster that has been discovered at an earlier time point. Cluster transitions provide insights about the nature of cluster changes: is a cluster a newly emerging cluster or a disappeared one or does some of its member move to different clusters.

Cluster transitions detection takes as input two sets of document clustering results at two consecutive time points. It finds relationships between all clusters in two clustering result. Each cluster is assumed to contain lists of document names and keywords.

To find the transitions between two sets of clustering results, for each cluster C_i in D_i and each cluster C_j in D_j, it measures the degree to which C_i overlaps C_j. We use the following function for this work:

$$overlap(C_i, C_j) = |C_i \cap C_j| / |C_i|, \quad (1)$$

where $|C_i|$ and $|C_j|$ are the number of documents in C_i and C_j, respectively. It uses threshold which allows zero to multiple match and returns lists of clusters in D_j which is either a match list or a split list of clusters in D_i.

3.2 Keyword search heuristic

A user can issue a query to the system. A query consists of a set of one or more keywords. This approach assumes that each cluster has associated lists of keywords and keyword scores. It looks for matched clusters corresponding to the query in the cluster transitions.

3.2.1 Matched cluster candidate detection

All clusters which match a given query are detected. A query Q, for example, consists of three keywords, $Q = \{t_1, t_2, t_3\}$. The "OR" boolean operator is used as the search condition $\{t_1 \text{ OR } t_2 \text{ OR } t_3\}$.

For each cluster in the clustering results, compare the query against the keyword list of the cluster. If the query is found in a cluster and the score is larger than a threshold, store the cluster in the matched cluster candidate list for the time point and the score of Q of the cluster. The score is used in summarization heuristic in Section 3.3.

3.2.2 Matched cluster candidate pruning using cluster transitions

The matched cluster candidate lists of the query at all time points are pruned based on cluster transitions. The intuition of pruning using cluster transitions is that cluster transitions convey topicwise relationship between clusters. They define the general semantic connection of the two clusters. If a query is found in two clusters of two consecutive time points, however there is no transition between the two clusters, the occurrence of the query in the two clusters is considered not topically related to each other. The query keyword may be a general term and may have been used just as a description of particular events.

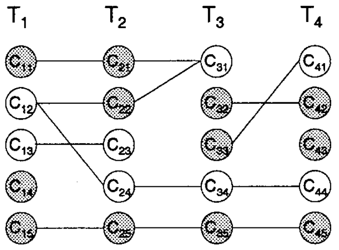

Figure 1: Cluster transitions and matched cluster candidates of query Q

Each matched cluster candidate list is mapped to the cluster transitions.

- If the cluster candidates scatter in the transition, i.e., they do not locate in any transition paths. The query keywords of the search condition are considered generic and not topic specific. Therefore no summarized descriptive sentences can be generated.
- When the cluster candidates nearly fit some transition paths, prune those clusters which do not locate in transition paths off the candidate list. If there are more than one transition path started from the same period fitted the cluster candidates, the transition path with the most number of cluster candidates is selected. All cluster candidates on the transition path and their scores and time points are chosen as the result to the query. Consider, for example, matched cluster candidates of query Q marked in lighter color in Fig. 1. C_{31} and C_{41} are pruned off the candidate list and $\{C_{12}, C_{24}, C_{34}, C_{44}\}$ are selected as the result to the query.

3.3 Summarization heuristic

Given result clusters for a query and their scores and time points, we uses connection rules and templates to summarize the result and generate high level descriptive sentences about the query.

3.3.1 Connection rules

Consider the example in Section 3.2.2, we obtained the following result $R = \{C_{12}, C_{24}, C_{34}, C_{44}\}$ for the query $Q=$ {"Asashoryu", "soccer"}, and the scores of query keywords in each cluster are as follows: C_{12}["Asashoryu"⇒ 0.8, "soccer"⇒ 0.6] at T_1, C_{24}["Asashoryu"⇒ 0.7, "soccer"⇒ 0.4] at T_2, C_{34}["Asashoryu"⇒ 0.5] at T_3, and C_{44}["soccer"⇒ 0.2] at T_4, where T_1 = Jan 1, T_2 = Jan 2, T_3 = Jan 3, and T_4 = Jan 4. The following connection rules are used to join the results.

- R1: Merge the time point of the same instance of keywords. For example, {"Asashoryu", "soccer"} ⇒ {Jan 1, Jan 2}, {"Asashoryu"} ⇒ {Jan 3}, {"soccer"} ⇒ {Jan 4}.
- R2: Average the score of instances of more than one keyword. It is used to determine "hotness" of the query at the time point. For example, {"Asashoryu", "soccer"} ⇒ {0.7, Jan 1}, {"Asashoryu", "soccer"} ⇒ {0.55, Jan 2}.

3.3.2 Templates

The following templates are used to generate descriptive sentences after connection rules have been applied on the results.

- T1: {Query} became hot on {time}.
- T2: {Query} appeared on {time}.
- T3: {Query} re-appeared on {time}.
- T4: T2 and T3.
- T5: T2 and T1.
- T6: T4 and T1.

Definitions of templates: "hot" means the average score of a query is large to a certain extent. Other templates should be self explanatory. We assume three degrees of "hotness".

- If score is $< a$, it is "not hot". "not hot" is not reflected in the descriptive sentences.
- If score is $\geq a$ and $< b$, it is "fairly hot".
- If score is $\geq b$ and $< c$, it is "rather hot".
- If score is $\geq c$, it is "very hot".

The values of parameters a, b and c can vary depending on the target data.

By setting $a = 0.4$, $b = 0.6$, and $c = 0.8$, the descriptive sentences for $Q=$ {"Asashoryu", "soccer"} from the above example are: {"Asashoryu", "soccer"} appeared on {Jan 1, Jan 2} and became rather hot on {Jan 1} and fairly hot on {Jan 2}. {"Asashoryu"} appeared and became fairly hot on {Jan 3}. {"soccer"} appeared on {Jan 4}.

4 Conclusion

In this paper, we have described the framework to generate descriptive sentences as overall summary from the whole collection of document clustering results using cluster transitions, given a query. The framework includes the detection of cluster transitions, keyword search from the cluster transitions and generation of descriptive sentences. Our future work is to crystallize the proposed framework and to implement it.

Acknowledgements

This research is partly supported by the Grant-in-Aid for Scientific Research (19300027) from Japan Society for the Promotion of Science (JSPS) and the Grant-in-Aid for Scientific Research (19024006) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. In addition, this work is supported by the grants from Hoso Bunka Foundation.

References

[1] Allan, J. (ed.): Topic Detection and Tracking: Event-based Information Organization. Kluwer, Boston (2002)

[2] Baeza-Yates, R., and Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Harlow, England (1999)

[3] Ishikawa, Y., Hasegawa, M.: T-Scroll: Visualizing Trends in a Time-series of Documents for Interactive User Exploration. In: Proc. of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), pp. 235–246 (2007)

[4] Khy, S., Ishikawa, Y., Kitagawa, H.: A Novelty-based Clustering Method for On-line Documents, World Wide Web Journal, DOI 10.1007/s11280-007-0018-9

[5] Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., Schult, R.: MONIC – Modeling and Monitoring Cluster Transitions. In: Proc. of KDD Conference, pp. 706–711 (2006)

[6] Yang, Y., Carbonell, J.G., Brown, R.G., Pierce, T., Archibald, B.T., Liu, X.: Learning Approaches for Detecting and Tracking News Event. IEEE Intel. Sys. **14**(4), July/August, pp. 32–43 (1999)