Rabia IRFAN, Sharifullah KHAN, Kashif RAJPOOT, Ali Mustafa QAMAR, 2018. TIE algorithm: a layer over clustering-based taxonomy generation for handling evolving data. *Frontiers of Information Technology & Electronic Engineering*, 19(6):763-782. https://doi.org/10.1631/FITEE.1700517

TIE algorithm: a layer over clustering-based taxonomy generation for handling evolving data

Key words: Taxonomy; Clustering algorithms; Information science; Knowledge management; Machine learning

Corresponding author: Rabia IRFAN
E-mail: 12phdrirfan@seecs.edu.pk

ORCID: https://orcid.org/0000-0002-7789-5338

Motivations

1. Taxonomy is a hierarchical organization of data and is an effective mean of categorizing and organizing the data.

2. Manual generation of taxonomy is tedious and cumbersome, so many automatic taxonomy generation techniques exist.

3. Data frequently evolves in today's digitally connected world.

4. However, majority of the existing taxonomy generation techniques does not pay much attention to this nature of data and the generated taxonomy remains static even if data evolves.

5. Taxonomy generated from this data should evolve with evolving data.

Main ideas

1. Taxonomy can be evolved in two ways:

   (1) Regeneration of new taxonomy from scratch;

   (2) Incremental evolution of existing taxonomy .

2. Regeneration approach has some limitations as:

   (1) Regeneration does not periodically occur, so taxonomy is not an accurate representation of underlying data most of the time.

   (2) Regeneration starts the whole process from scratch, so it is time consuming and cost inefficient.

3. Focus of this research is on the evolution of existing taxonomy incrementally, as the underlying data evolves.

4. This work presents TIE algorithm which is a novel algorithm for evolving taxonomy incrementally whenever changes in data occur.

Methods

1. TIE uses a clustering-based technique for the generation of initial taxonomy.

2. TIE takes new document, existing taxonomy, and its respective hierarchical structure as inputs, and its process revolves around the adjustment of new documents in the hierarchical structure and ultimately in the existing taxonomy

3. It comprises two steps:

   (1) Identifying the closest cluster: This step identifies the closest cluster as a possible candidate for a new document to be adjusted in.

   (2) Reorganizing the existing taxonomy: Reorganizing clusters in the hierarchy in case new documents effect quality of clusters in the respective hierarchical structure.

Major results

Time-based evaluation

1. Table 5 in case of HAC and Table 8 in case of bisect *K*-means, clearly reflect the essence of incremental evolution of taxonomy in comparison to regeneration. As expected, the amount of time it takes to incrementally evolve is less than that of regeneration.

Table 5 Results of time-based evaluation in case of HAC

With HAC	T_{run} ($\pm\sigma$) (min)	
	TGP	TIE
200	22.842 (±0.308)	–
200+20=220	24.725 (±0.080)	4.905 (±0.232)
220+30=250	28.030 (±0.170)	7.182 (±0.245)
250+40=290	46.760 (±0.009)	9.810 (±0.180)
290+50=340	59.390 (±0.140)	15.601 (±0.238)
340+60=400	70.088 (±0.437)	20.675 (±0.337)

Table 8 Results of time-based evaluation in case of bisect *K*-means

Data set	t_{run} ($\pm\sigma$) (min)	
	TGP	TIE
200	21.157 (±0.012)	–
200+20=220	22.569 (±0.310)	4.438 (±0.010)
220+30=250	23.981 (±0.048)	6.479 (±0.011)
250+40=290	30.642 (±0.211)	10.911 (±0.020)
290+50=340	45.093 (±0.224)	13.714 (±0.014)
340+60=400	57.503 (±0.126)	19.704 (±0.050)

Major results

Quality-based evaluation (lexical)

2. Lexical quality of taxonomy produced as a result of regeneration is slightly better than the one obtained as a result of evolution, shown in Tables 6 (HAC) and 9 (bisect *K*-means).

Table 6 Results of lexical quality based evaluation in case of HAC

Data set	LP (average±σ)		LR (average±σ)		LF	
	TGP	TIE	TGP	TIE	TGP	TIE
200	0.580±0.010	–	0.457±0.004	–	0.511	–
200+20=220	0.581±0.020	0.522±0.008	0.440±0.001	0.425±0.013	0.501	0.469
220+30=250	0.539±0.017	0.516±0.010	0.434±0.004	0.415±0.035	0.481	0.460
250+40=290	0.532±0.009	0.501±0.065	0.424±0.002	0.402±0.056	0.472	0.446
290+50=340	0.512±0.042	0.499±0.061	0.409±0.002	0.392±0.014	0.455	0.439
340+60=400	0.507±0.004	0.490±0.061	0.411±0.003	0.401±0.045	0.454	0.441

Table 9 Results of lexical quality-based evaluation in case of bisect *K*-means

Data set	LP (average±σ)		LR (average±σ)		LF	
	TGP	TIE	TGP	TIE	TGP	TIE
200	0.445±0.007	–	0.323±0.002	–	0.374	–
200+20=220	0.405±0.009	0.380±0.015	0.338±0.086	0.327±0.020	0.369	0.352
220+30=250	0.402±0.010	0.354±0.012	0.351±0.005	0.323±0.002	0.375	0.338
250+40=290	0.408±0.011	0.358±0.001	0.317±0.003	0.292±0.016	0.356	0.322
290+50=340	0.395±0.012	0.354±0.012	0.311±0.094	0.310±0.002	0.347	0.331
340+60=400	0.390±0.037	0.338±0.004	0.304±0.083	0.296±0.111	0.342	0.316

Major results

Quality-based evaluation (Hierarchical)

3. Hierarchical quality of taxonomy produced as a result of regeneration is slightly better than the one obtained as a result of evolution, shown in Tables 7 (HAC) and 10 (bisect *K*-means).

Table 7 Results of hierarchical quality based evaluation in case of HAC

Data set	HP (average±σ)		HR (average±σ)		HF	
	TGP	TIE	TGP	TIE	TGP	TIE
200	0.351±0.099	–	0.330±0.010	–	0.340	–
200+20=220	0.341±0.007	0.338±0.006	0.323±0.010	0.283±0.001	0.331	0.308
220+30=250	0.346±0.068	0.307±0.004	0.319±0.002	0.275±0.013	0.332	0.290
250+40=290	0.326±0.004	0.304±0.020	0.298±0.033	0.277±0.003	0.311	0.292
290+50=340	0.312±0.003	0.295±0.003	0.295±0.003	0.276±0.004	0.304	0.285
340+60=400	0.305±0.009	0.291±0.086	0.286±0.006	0.264±0.003	0.295	0.277

Table 10 Results of hierarchical quality-based evaluation in case of bisect *K*-means

Data set	HP (average±σ)		HR (average±σ)		LF	
	TGP	TIE	TGP	TIE	TGP	TIE
200	0.244±0.089	–	0.217±0.054	–	0.229	–
200+20=220	0.241±0.011	0.217±0.013	0.228±0.032	0.198±0.011	0.234	0.207
220+30=250	0.242±0.008	0.202±0.008	0.220±0.051	0.177±0.052	0.231	0.189
250+40=290	0.233±0.002	0.197±0.004	0.204±0.076	0.203±0.095	0.218	0.199
290+50=340	0.218±0.031	0.196±0.002	0.193±0.007	0.185±0.065	0.205	0.190
340+60=400	0.208±0.053	0.179±0.014	0.188±0.006	0.162±0.034	0.197	0.170

Major results

Quality-time ratio

4. We can see from all the graphs shown in Figs. 2 and 3 that r_{QT} is higher in the case of evolution for datasets of different sizes.

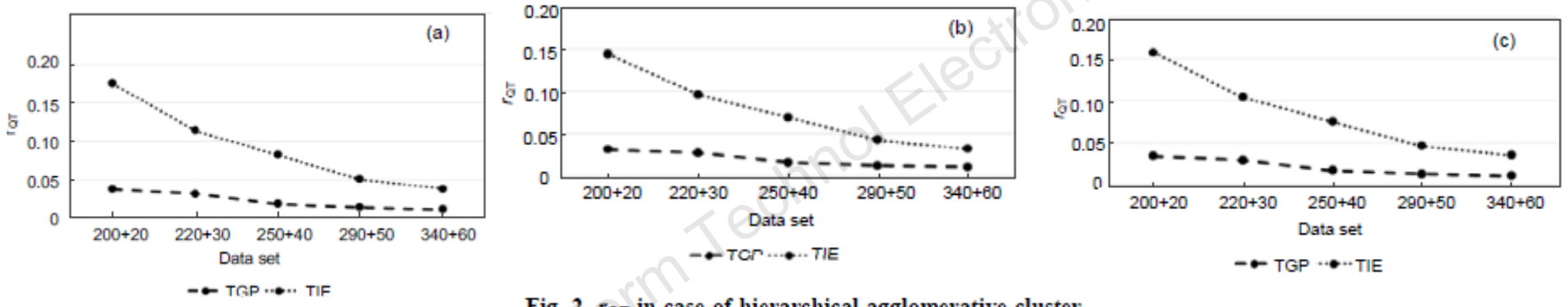

Fig. 2 r_{QT} in case of hierarchical agglomerative clustering: (a) LP+HP; (b) LR+HR; (c) LF+HF

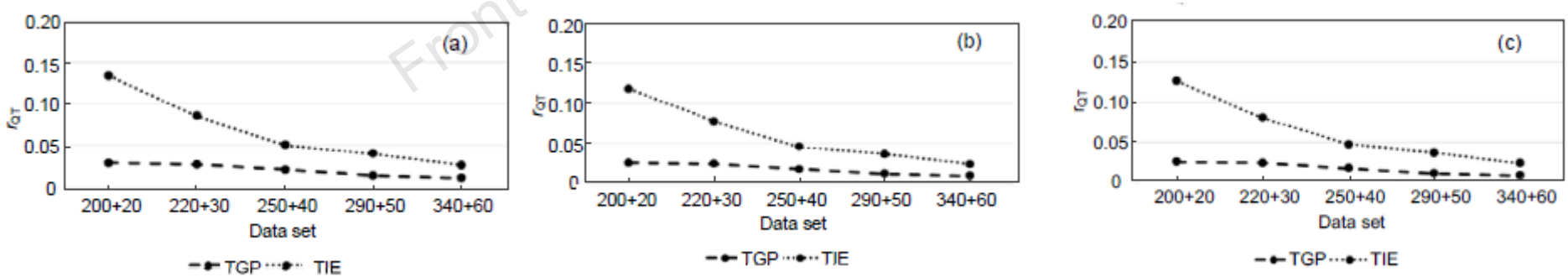

Fig. 3 r_{QT} in case of bisect K-means: (a) LP+HP; (b) LR+HR; (c) LF+HF

Conclusions

1. In this paper, a TIE algorithm is designed to incrementally evolve an existing taxonomy to adjust changes that occur in underlying data.

2. The algorithm is compared with a taxonomy regeneration approach based on complexity analysis and empirical evaluation.

3. The complexity analysis of the algorithm demonstrates that it is better than regeneration in terms of time, and that it is independent of the underlying clustering approach used for taxonomy generation.

4. The empirical evaluation is performed using a dataset of scholarly articles selected from the computing domain, based on three parameters: time-based, quality-based, and quality-time based.

Conclusions

5. Time-based evaluation clearly shows that TIE algorithm takes less time to adjust new documents in an existing taxonomy.

6. Although taxonomy regeneration shows better results in terms of quality, and the quality-time ratio of the TIE algorithm indicates that the rate of improvement in taxonomy quality per unit time is better than that of regeneration.

7. The results of sensitivity analysis of the TIE algorithm show that it performs better when the new data arrives in small chunks.

8. In future, the proposed algorithm can be applied to discover emerging trends and patterns in social media where data is rapidly evolving.