Mining Temporal Evolution of Criminal Behaviors

Gianvito Pio   Michelangelo Ceci   Donato Malerba

University of Bari “Aldo Moro”
Department of Computer Science - Via Orabona, 4 - 70125 Bari, Italy

20th Italian Symposium on Advanced Database Systems (SEBD 2012)

Outline

Risk Identification and Analysis

Risk Identification and Analysis: investigation activities with the goal of defending a Nation or a community against potential threats.

Studies in the literature [Chen et al., 2004, Jonas and Harper, 2006, Seifert, 2010] have proved the effectiveness of Data Mining techniques in supporting the investigative activity in risk identification and analysis.

Orthogonally: Topic Detection and Tracking

Over the last years, **Topic Detection and Tracking (TDT)** [Allan, 2002 Yang et al., 1999, Brants et al., 2003] is being recognized as an important research area in Data Mining.

Research lines in TDT ([Chung and Mcleod, 2005]):

- News segmentation
- New topic detection
- Topic tracking

Exploiting Topic Tracking techniques

- **Idea:** to exploit topic tracking techniques in the risk identification and analysis
- **Goal:** to discover evolutions of criminal behaviors over time
- **Input:** streams of time-stamped news (or, generally, documents) associated to criminals
- **Method:** incremental analysis of streams of news in order to identify clusters of similar criminals and represent their evolution over time

Related work...

... in cluster evolution analysis for topic tracking

- **[Leskovec et al., 2009, Zhu and Shasha, 2003]:** tracking topics, ideas and "memes" from news
- **[Kleinberg, 2002, Aggarwal, 2005]:** an evolution is discovered when a particular data mining model becomes stale because of the underlying change in the data distribution
- **[Zhong, 2005]:** incremental and neural network based k-means applied to news (incremental update of the closest cluster)
- **[Agarwal et al., 2010]:** clustering of manually labeled blogs (generation of a so called "collective wisdom")
- **[Li et al., 2009]** clustering stories into topics from different blogs. Two-phases clustering (initial static step and incremental distance-based update of clusters)

Related work...

... in risk identification and analysis

- **[Chen et al., 2004]**: different algorithms (for clustering, classification, social network analysis, etc.) are proposed for analyzing data about criminal activities (e.g. money laundering identification, criminal profiling, etc.)
- **[Schroeder et al., 2007, Ozgul et al., 2007]**: social/criminal network link analysis
- **[Chen et al., 2004, Chau et al., 2002, Xu and Chen, 2004]**: extraction of crime entity associations from textual documents

Main differences

Our approach...

- does not consider variations and evolution (in volume) of short and distinctive phrases in the news, but the evolution of each single criminal to which multiple news can be associated
  - $\rightarrow$ **Unit of analysis:** criminal
- discovers evolutions expressed according to the relevant terms that allow us the representation and characterization of criminals

TB-CREDIS

The framework **TB-CREDIS (Time-Based CRiminal Evolution DIScoverer)** consists of the following phases:

- partitioning the whole time period of analysis in disjoint, adjacent and equal-size time intervals (*time-windows*);
- VSM representation of the all the time-stamped documents, which are implicitly associated to a time-window;
  - feature selection;
- identification of the semantic position of each criminal in each time window;
- clustering of criminals for each time window;
- evolution discovery and analysis.

Feature Selection...

- The number of terms extracted from documents collection is usually high
  - $\rightarrow$ a feature selection phase is necessary
- No additional information to guide the feature selection (e.g. target attribute)
  - $\rightarrow$ unsupervised feature selection

Feature Selection - Variance

Variance: selects the top-k terms with the highest variance value

$$Score(t_r) = \frac{1}{n-1}\sum_{j=1}^{n}(s_{r,j} - \bar{s}_r)^2$$

$\mathbf{s_{r,j}}$ = weight of the term t_r in the document d_j

$\mathbf{\bar{s}_r}$ = average weight of the term t_r in the whole documents collection

Feature Selection - Variance

Variance: selects the top-*k* terms with the highest variance value

$$Score(t_r) = \frac{1}{n-1}\sum_{j=1}^{n}(s_{r,j} - \bar{s}_r)^2$$

Strong points:

- It selects terms which well discriminate between documents
- Low time complexity

Weak points:

- It does not take into account the correlation between selected terms
- Selected terms may not preserve the similarity/dissimilarity between documents

Feature Selection: MIGRAL-CP

MIGRAL-CP (MInimum GRAph Loss with Correlation Penalty): selects the top uncorrelated k terms which best preserve the similarity/dissimilarity between documents :

$$Score_1(t_r) = \frac{1}{2}\left(1 - \frac{1}{n}\sum_{j=1}^{n} \rho(V_j, F_{r,j})\right)$$

where:

- $V_j = [v_{j,1}, v_{j,2}, \dots, v_{j,n}]$ are the dissimilarity values between the document j and all the other documents, using all the terms (Gaussian distance on TF representations)
- $F_{r,j} = \left[(s_{r,j} - s_{r,1})^2, (s_{r,j} - s_{r,2})^2, \dots, (s_{r,j} - s_{r,n})^2\right]$ are the dissimilarities between the document j and all the other documents, using the term t_r only
- ρ is the Pearson correlation coefficient

Feature Selection: MIGRAL-CP

$$Score_i(t_r) = Score_{i-1}(t_r) \times (1 - penalty(t_r, \hat{t}_{i-1}))$$

At each iteration i, scores are reduced according to a penalty function which considers the correlation between the term t_r and the term that has been selected in the previous iteration ($\hat{t}_{i-1}$)

$\rightarrow$ **prevents the selection of redundant features**

We use: $penalty(t_r, \hat{t}_{i-1}) = \max\left(0, \left|\rho(t_r, \hat{t}_{i-1})\right| - \gamma\right)$,
where $0 \leq \gamma \leq 1$.

Representing criminals

Criminal: a point in the h-dimensional space, which better represents his/her semantic position (crime typologies).

The semantic position of each criminal c is identified:

- for each time window τ_z
- according to the set of documents he/she is associated to, in the considered time window
- (possibly) considering documents belonging to previous time-windows

Example ($h = 7$): [attack: 0.593; fire: 0.371; claim: 0.271; suspect: 0.1; report: 0.057; injur: 0.057; islam: 0.05]

Representing criminals: Time-weighted centroid

$$X(c, \tau_z, h) = \frac{\sum_{<d_j,\tau_j> \in S_{c,\tau_z,h}} p_{\tau_z,\tau_j}(h) \times w_{d_j}}{\sum_{<d_j,\tau_j> \in S_{c,\tau_z,h}} p_{\tau_z,\tau_j}(h)},$$

where:

- $S_{c,\tau_z,h}$ is the set of documents associated to the criminal c, belonging to the considered time window τ_z or one of the previous $h-1$ time windows
- $p_{\tau_z,\tau_j}(h) = 1 - \frac{z-j+1}{h}$ is the time fading-factor which reduces the effect of the document d_j according to the distance between the considered time window (τ_z) and the time window τ_j the considered document is associated to.

Representing criminals: Max Density Point

Each document is replaced by a *k*-dimensional Gaussian function:

$$d_j'(x) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - s_{i,j})^2}{2\sigma^2}}$$

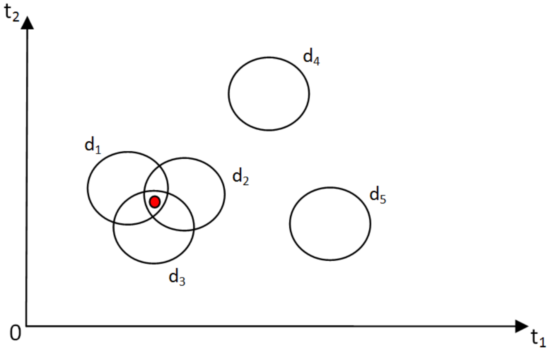
where $\sigma \in [0,1]$ is a parameter that defines the width of the Gaussian function.

The criminal position is that which presents the maximum value of the sum of time-weighted Gaussian functions associated to documents:

$$X(c, \tau_z, h) = \arg\max_{x \in [0,1]^k} \sum_{<d_j, \tau_j> \in S_{c,\tau_z,h}} p_{\tau_z,\tau_j}(h) \times d_j'(x)$$

Representing criminals: Max Density Point

An example of documents represented in a 2-dimensional space (top view). The red point represents the identified semantic position of the criminal.



Representing criminals: Max Density Point

Computational optimization:

- equal-width discretization of the space $[0,1]^k$ into Φ^k,
  where $\Phi = \left\{0, \frac{1}{\beta}, \frac{2}{\beta}, \ldots, \frac{\beta-1}{\beta}, 1\right\}$
- greedy search, focusing only on the points for which the $d'_j(\cdot)$ functions reach the maximum values
  - $\rightarrow$ being y the value in which the Gaussian function assumes the maximum value on a dimension, we search in $[y-\sigma; y+\sigma]$
- the criminal position at the time-window τ_z can only be the position at the previous time-window or around new documents (belonging to τ_z)
  - $\rightarrow$ the search can be limited to the areas interested by the documents belonging to τ_z and to $X(c, \tau_z, h)$
- parallel computation on multiple CPUs

Clusters evolution discovering

Observations on the clustering step:

- there is no guarantee that all the crime typologies are present in each time window
- there is no way to know a-priori the real number of crime typologies for each time-window

Proposed solution:

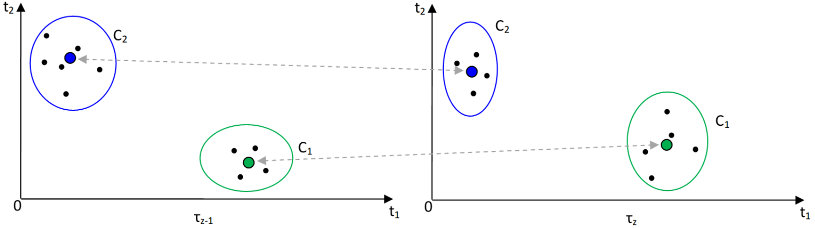
- K-Means clustering algorithm
- automatic estimation of the most appropriate number of clusters, using Principal Component Analysis (PCA) [Jolliffe, 2002], for each time-window

Clusters evolution discovering

Once clustering is performed for each time-window, it is possible to identify:

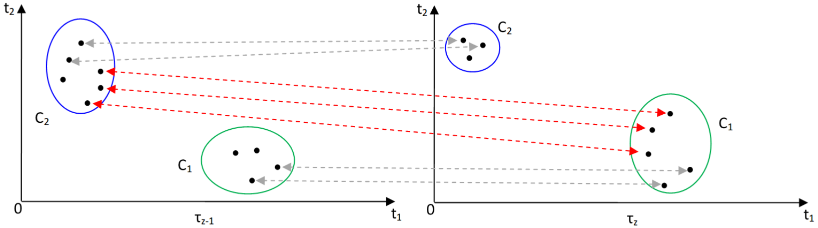
- **the position** of each cluster in the terms space. Analyzing the terms with the highest weight in the cluster can give an idea about the crime typology it represents
- **a matching** between clusters of different time windows according to the similarity between the clusters' centroids
- **the number of criminals** which have evolved from the crime typology represented by two different clusters belonging to two adjacent time-windows

Clusters evolution discovering



An example of matching found between two clusters belonging to different time windows, analyzing the centroids' similarity.

Clusters evolution discovering



An Example of a discovered criminal evolution. Three criminals have moved from the cluster C_2 in τ_{z-1} to the cluster C_1 in τ_z.

Experiments

Datasets:

- Synthetic dataset
- Global Terrorism Database (GTD)

Evaluation:

- average Q-Modularity [Newman, 2006] of the obtained clustering
- running time

Experiments: Synthetic data

Synthetic dataset characteristics:

- 10 consecutive annual time windows (from 2001 to 2010)
- 100 criminals
- up to 200 documents for each criminal
- 7 crime typologies generated from 7 specific vocabularies and a generic English vocabulary (noise terms)
- each criminal has the 30% of probability to change crime typology

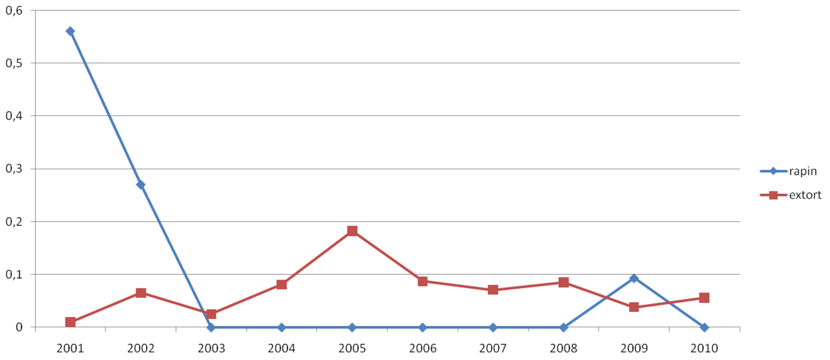
Experimental setup:

- Feature selection: $k = 10$, $\gamma = 0.5$ (MIGRAL-CP)
- Max Density Point method: $\beta = 20$
- Variable h, σ and variance (for PCA)

Experiments: Synthetic data

				Variance		MIGRAL-CP	
Position	**h**	σ	**Var**	**time**	**q-mod**	**time**	**q-mod**
Centroid	2	-	80%	00:20:52	0.157	00:55:54	0.198
Centroid	2	-	90%	00:20:52	0.150	00:55:54	0.209
Centroid	2	-	80%	00:20:58	0.102	00:55:59	0.142
Centroid	2	-	90%	00:20:58	0.101	00:55:59	0.143
Centroid	2	-	80%	00:21:00	0.080	00:56:01	0.114
Centroid	2	-	90%	00:21:00	0.081	00:56:01	0.115
MaxDensity	2	0.05	80%	00:21:47	0.322	00:56:44	0.392
MaxDensity	2	0.05	90%	00:21:47	0.356	00:56:44	0.380
MaxDensity	2	0.10	80%	01:20:35	0.335	01:50:08	0.375
MaxDensity	2	0.10	90%	01:20:35	0.357	01:50:08	0.373
MaxDensity	5	0.05	80%	00:20:19	0.341	00:57:12	0.399
MaxDensity	5	0.05	90%	00:20:19	0.365	00:57:12	0.379
MaxDensity	5	0.10	80%	01:53:13	0.363	02:19:22	0.350
MaxDensity	5	0.10	90%	01:53:13	0.366	02:19:22	0.368
MaxDensity	10	0.05	80%	00:22:27	0.339	00:57:28	0.386
MaxDensity	10	0.05	90%	00:22:27	0.385	00:57:28	0.371
MaxDensity	10	0.10	80%	02:10:17	0.369	02:35:54	0.354
MaxDensity	10	0.10	90%	02:10:17	0.372	02:35:54	0.369

Experiments: Synthetic data



A cluster evolution in the synthetic dataset. TF-IDF values are plotted.

Experiments: Real data

Read dataset characteristics:

- Global Terrorism Database (GTD)[1]
- Information on about 98,000 terrorism events (1970-2010)
- 13 annual time-windows have been considered (1998-2010)
- A total of 11,225 news about 82 criminals/organizations

Experimental setup:

- Feature selection: $k = 15$, $\gamma = 0.5$ (MIGRAL-CP)
- Max Density Point method: $\beta = 20$
- $\sigma = 0.05$ and variable h and variance (for PCA)

[1] http://www.start.umd.edu/gtd/

Experiments: Real data

				Variance		MIGRAL-CP	
Position	**h**	σ	**Var**	**time**	**q-mod**	**time**	**q-mod**
Centroid	2	-	90%	00:09:01	0.294	39:54:44	0.245
Centroid	2	-	95%	00:09:01	0.319	39:54:44	0.270
Centroid	5	-	90%	00:09:06	0.297	39:54:48	0.224
Centroid	5	-	95%	00:09:06	0.316	39:54:48	0.249
Centroid	10	-	90%	00:09:10	0.304	39:54:51	0.232
Centroid	10	-	95%	00:09:10	0.322	39:54:51	0.245
MaxDensity	2	0.05	90%	110:41:36	0.322	100:17:46	0.447
MaxDensity	2	0.05	95%	110:41:36	0.509	100:17:46	0.479
MaxDensity	5	0.05	90%	137:41:36	0.325	118:59:17	0.454
MaxDensity	5	0.05	95%	137:41:36	0.521	118:59:17	0.487
MaxDensity	10	0.05	90%	144:20:21	0.400	126:06:27	0.452
MaxDensity	10	0.05	95%	144:20:21	0.524	126:06:27	0.479

Considerations

- the **MIGRAL-CP algorithm** leads to higher clustering quality in the synthetic dataset, at the price of significantly higher running times
- the **Max Density Point method** always significantly outperforms the centroid method on both datasets, at the price of slightly higher running times
- the best combination appears to be **MaxDensity-Variance** in the case of a relatively small number of clusters (Var=90%)

Conclusions

A framework which is able to incrementally extract knowledge from time-stamped news has been proposed.

Three sequential steps:

- VSM representation of documents (feature selection)
- identification of the semantic position of subjects
- clustering and evolution analysis

Evaluation has been conducted in the context of risk identification and analysis in order to understand the evolution of criminal behaviors.

Future work

- Analytic identification of the value of σ, with respect to h, such that the global optimum is guaranteed
- A detailed qualitative evaluation on the evolutions discovered on real datasets
- An analysis of the effects of different size of time-windows on the obtained results

Thank you for your attention
Questions?

Advertisement: Workshop NFMCP @ ECMLPKDD2012

NFMCP: New Frontiers in Mining Complex Patterns

Annalisa Appice, Michelangelo Ceci, Corrado Loglisci, Giuseppe Manco, Elio Masciari and Zbigniew Ras

Important Dates
- Paper submission: Friday, June 29, 2012
- Acceptance notification: Friday July 20, 2012
- Camera-ready of accepted papers: Friday August 3, 2012

- Agarwal, N., Galan, M., Liu, H., and Subramanya, S. (2010).

  Wiscoll: Collective wisdom based blog clustering.
  Inf. Sci., 180:39–61.
- Aggarwal, C. C. (2005).
  On change diagnosis in evolving data streams.
  IEEE Trans. Knowl. Data Eng., 17(5):587–600.
- Allan, J., editor (2002).
  Topic Detection and Tracking: Event-based Information Organization.
  Kluwer International Series on Information Retrieval. Kluwer.
- Brants, T., Chen, F., and Farahat, A. (2003).
  A system for new event detection.

In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 330–337. ACM.

- Chau, M., Xu, J. J., and Chen, H. (2002).
  Extracting meaningful entities from police narrative reports.
  In *Proc. of the national conference on Digital government research*, dg.o '02, pages 1–5. Digital Government Society of North America.

- Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., and Chau, M. (2004).
  Crime data mining: A general framework and some examples.
  IEEE Computer, 37:50–56.

- Chung, S. and Mcleod, D. (2005).

Dynamic Pattern Mining: An Incremental Data Clustering Approach.
pages 85–112.

Jolliffe, I. T. (2002).
Principal Component Analysis.
Springer, second edition.

Jonas, J. and Harper, J. (2006).
Effective Counterterrorism and the Limited Role of Predictive Data Mining.
Policy Analysis, 584.

Kleinberg, J. (2002).
Bursty and hierarchical structure in streams.

In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 91–101, New York, NY, USA. ACM.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009).
Meme-tracking and the dynamics of the news cycle.
In *KDD '09*, pages 497–506, New York, NY, USA. ACM.

Li, X., Yan, J., Fan, W., Liu, N., Yan, S., and Chen, Z. (2009).
An online blog reading system by topic clustering and personalized ranking.
ACM Trans. Internet Technol., 9:9:1–9:26.

Newman, M. E. J. (2006).
Modularity and community structure in networks.

Proceedings of the National Academy of Sciences, 103(23):8577–8582.

- Ozgul, F., Bondy, J., and Aksoy, H. (2007).
  Mining for offender group detection and story of a police operation.
  In *Proceedings of the sixth Australasian conference on Data mining and analytics - Volume 70*, AusDM '07, pages 189–193, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

- Schroeder, J., Xu, J., Chen, H., and Chau, M. (2007).
  Automated criminal link analysis based on domain knowledge: Research articles.
  J. Am. Soc. Inf. Sci. Technol., 58.

- Seifert, J. W. (2010).

Data Mining and Homeland Security: An Overview.
Bibliographisches Institut AG, Mannheim, W. Germany, Germany.

- Xu, J. J. and Chen, H. (2004).
  Fighting organized crime: Using shortest-path algorithms to identify associations in criminal networks.
  Decis. Support Syst., 38:473–487.

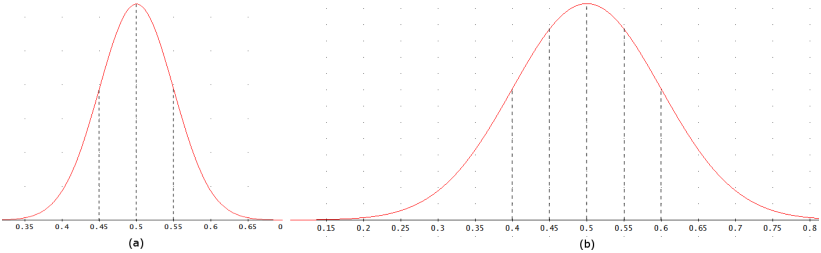
- Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., and Liu, X. (1999).
  Learning approaches for detecting and tracking news events.
  Intelligent Systems and their Applications, IEEE, 14(4):32–43.

- Zhong, S. (2005).

Efficient streaming text clustering.
Neural Networks, 18(5-6).

Zhu, Y. and Shasha, D. (2003).
Efficient elastic burst detection in data streams.
In *ACM SIGKDD*, KDD '03, pages 336–345, New York, NY, USA. ACM.

Representing criminals: Max Density Point



A Gaussian function defined on a single dimension with $y = 0.5$, $\beta = 20$, $\sigma = 0.05$ (a) and $\sigma = 0.10$ (b). In (a) it would be enough to analyze only the values 0.45, 0.50 and 0.55, whereas in (b) it would be necessary to analyze also the values 0.40 and 0.60.