Detección de los patrones raros en un conjunto de datos semiestructurados

M. Montes-y-Gómez [1], **A. Gelbukh** [1] y **A. López-López** [2]

[1] Centro de Investigación en Computación (CIC-IPN), México.
mmontesg@susu.inaoep.mx, gelbukh@cic.ipn.mx

[2] Instituto Nacional de Astrofísica, Optica y Electrónica (INAOE), México.
allopez@inaoep.mx

Resumen. La detección de desviaciones es un problema importante de la minería de datos. Consideramos el descubrimiento de las desviaciones en un conjunto de datos semiestructurados; en particular, en un conjunto de datos representados con grafos conceptuales. Típicamente, la detección automática de desviaciones se realiza aplicando métodos estadísticos o basados en distancia. A diferencia de ellos, nuestro método se basa en las regularidades, es decir, características comunes de varios grafos de la colección. El método descubre los grafos que no comparten ninguna de estas regularidades. Además, se detectan los patrones raros en el conjunto, lo que es una manera resumida de expresar las desviaciones encontradas. Estos patrones se descubren a través del agrupamiento conceptual de los grafos. El método permite considerar las desviaciones con respecto a diferentes contextos (subconjuntos) elegidos por el usuario, es decir, desde diferentes perspectivas y a diferentes niveles de generalización.

Palabras clave: minería de datos, minería de texto, grafos conceptuales, desviaciones, agrupamiento.

1 Introducción

Actualmente, debido al vertiginoso avance científico y tecnológico de los últimos años, las instituciones tienen grandes capacidades de crear, almacenar y distribuir sus datos. Esta situación, entre otras cosas, ha aumentado la necesidad de nuevas herramientas que auxilien en la transformación de esta vasta cantidad de datos en información útil o nuevos conocimientos.

Un ejemplo de este tipo de herramientas son los sistemas de *minería de datos.* Estos sistemas permiten analizar grandes bases de datos y descubrir en ellas algunos patrones interesantes. Sin embargo, debido a su enfoque en el análisis de la información contenida en una base de datos, los sistemas de minería de datos no son apropiados para el análisis de otros tipos de información menos estructurada, como por ejemplo, la contenida en una colección de textos.

Este artículo se relaciona con este problema. En él se describe un método para el análisis automático de un conjunto de objetos complejos –posiblemente textos– representados por medio de *grafos conceptuales* (Sowa, 1984; Sowa, 1999).

Existen varios métodos para el análisis de un conjunto de grafos conceptuales. Algunos de ellos consideran su comparación (Myaeng and López-López, 1992; Mugnier and Chein, 1992; Mugnier, 1995; Montes-y-Gómez *et al*., 2000b; Montes-y-Gómez *et al*., 2001a), otros su aplicación en la recuperación de información (Myaeng, 1992; Ellis and Lehmann, 1994; Huibers *et al*., 1996; Genest and Chein, 1997; Montes-y-Gómez *et al*., 2000a), y otros mas su agrupamiento (Mineau and Godin, 1995; Godin *et al*., 1995; Bournaud and Ganascia, 1996; Bournaud and Ganascia, 1997; Montes-y-Gómez *et al*., 2001b).

En este artículo se introduce el problema de *detectar las desviaciones* (patrones raros) en un conjunto de grafos conceptuales. Básicamente se propone utilizar el agrupamiento conceptual de los grafos como un índice de la colección, y aprovechar esta estructura para localizar las desviaciones.

Esta estrategia de análisis no solo facilita la identificación de los grafos o patrones raros, sino también permite detectar desviaciones respecto a diferentes contextos (tipo especial de subconjunto) de la colección.

El resto del artículo se organiza de la siguiente manera. La sección 2 relata brevemente algunos trabajos relacionados. La sección 3 define el problema de detectar las desviaciones en un conjunto de grafos conceptuales. La sección 4 describe brevemente el tipo de agrupamiento de los grafos conceptuales que sirve como base para la detección de las desviaciones. La sección 5 presenta el método de detección de las desviaciones contextuales como tal. Finalmente, la sección 6 expone algunas conclusiones y discute los principales trabajos futuros.

2 Trabajos relacionados

Los métodos estadísticos tradicionales generalmente consideran que los datos raros (también llamados *desviaciones*) son una fuente de ruido, y por lo tanto, intentan minimizar sus efectos.

Diferente a este enfoque, algunos métodos de minería de datos se centran en la detección de estas desviaciones. Estos métodos consideran que las desviaciones pueden esconder conocimientos verdaderamente inesperados e interesantes.

Típicamente, los métodos para la detección de desviaciones emplean información adicional a los datos, por ejemplo: condiciones preestablecidas o restricciones de integridad (Guzmán, 1996). Solamente en algunas ocasiones estos métodos aprovechan la propia *redundancia* de los datos. Entre los métodos que aprovechan la redundancia de los datos destacan los siguientes tres enfoques (Han and Kamber, 2001):

- *Enfoque estadístico*. Este enfoque asume una distribución o modelo de probabilidad para los datos (por ejemplo, una distribución normal), y entonces, identifica las desviaciones con respecto a dicho modelo mediante una prueba de hipótesis.

  La aplicación de este modelo requiere que se conozcan la distribución de los datos, algunos de sus parámetros (media y varianza por lo regular) y el número de desviaciones esperadas. Una descripción amplia de este modelo se encuentra en (Barnett and Lewis, 1994).

- *Enfoque basado en distancia*. Este enfoque considera que el objeto *O* del conjunto *C* es una desviación con parámetros *p* y *d*, si al menos una fracción *p* de los objetos de *C* está a una distancia mayor que *d* de *O*. En otras palabras, el objeto *O* es una desviación si no tiene suficientes objetos "vecinos" (considerando una vecindad de radio *d*). En este caso, el usuario establece los parámetros *p* y *d*.

  Un algoritmo reciente para la detección de desviaciones basadas en distancia se describe en (Knorr and Ng, 1998).

- *Enfoque basado en regularidades*. Este enfoque asemeja la manera en que los humanos detectan las desviaciones (desde una perspectiva conceptual). Primero identifica las principales características –regularidades– del conjunto de objetos; después, con base en estas características, construye una descripción general de dicho conjunto; finalmente considera raros aquellos objetos que se "desvían" de la descripción general.

  Un método basado en este enfoque, aunque también un tanto estadístico, se presenta en (Arning *et al*., 1996). Básicamente, este método considera que las desviaciones son el conjunto de elementos que causan la mayor disimilitud en el conjunto de datos.

3 Desviaciones contextuales en grafos conceptuales

El método para la detección de las desviaciones contextuales en un conjunto de grafos conceptuales es, en términos generales, un método *basado en regularidades*. Este método se fundamenta en las siguientes consideraciones.

Dado un conjunto de grafos conceptuales $C = \{G_i\}$:

- Una *característica representativa* es cualquier generalización común g_c de más de m grafos del conjunto, siendo $m \geq f(n)$. En este caso n indica el número de grafos del conjunto y $f(n)$ es una función preestablecida por el usuario.

- Un *grafo conceptual raro* es un grafo que no tiene ninguna característica representativa[1], es decir, el grafo conceptual $G_r \in C$ es considerado raro, sí: $\nexists g_c : G_r < g_c$.

- Una *desviación* es un patrón descriptivo *d* de algunos grafos raros del conjunto *C*; en otras palabras una desviación es una expresión *resumida* de las rarezas del conjunto. Entonces, si se asume que *R* es el conjunto de grafos raros de *C* ($R \subseteq C$), la siguiente condición respecto a *d* se satisface: $\forall G_i \in C : G_i < d \Rightarrow G_i \in R$.

Así pues, dado un conjunto de grafos conceptuales $C = \{G_i\}$, donde cada grafo conceptual representa un texto diferente, una *desviación contextual* es una expresión de la forma: $g_i : g_j (r, s)$.

En esta expresión, g_i es el contexto y g_j es la descripción de algunos grafos *raros* en dicho contexto (de acuerdo con la función $f(n)$ preestablecida); *r* es el grado de rareza de

[1] Si no puede determinarse ninguna característica representativa del conjunto de grafos, entonces tampoco es posible (conceptualmente adecuado) detectar alguna desviación.

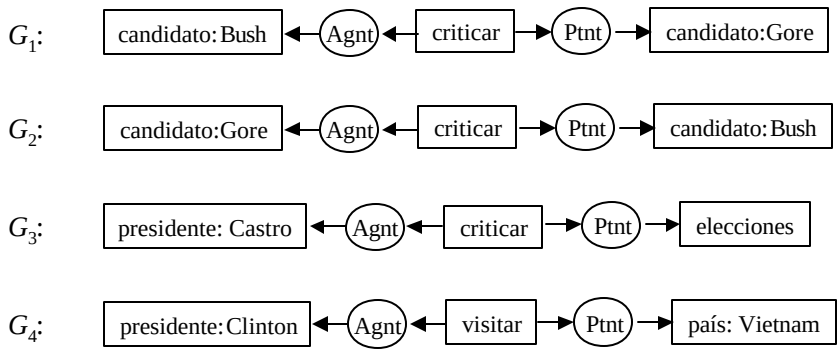

Figura 1. Un pequeño conjunto de grafos conceptuales.

los grafos conceptuales descritos por g_j en el contexto g_i, y s es el soporte de la desviación, es decir, es la representatividad del contexto g_j en el conjunto C.

Básicamente, esta expresión indica que: dentro del subconjunto de grafos conceptuales que contienen el grafo g_i, el cual representa el s% del conjunto total de grafos, todos los grafos que contienen el grafo g_j son *raros*; siendo éstos el c% de los grafos que contienen g_i.

Por ejemplo, la siguiente desviación contextual indica que 32% de los grafos de una colección hablan de animales, y que en este contexto sólo el 4% menciona un ave de rapiña.

[animal]: [ave]→ (tipo)→[rapiña] (32%,4%)

La detección de las desviaciones en un conjunto de grafos conceptuales se define como el problema de encontrar todas las desviaciones contextuales g_i: g_j (r/s) de acuerdo con una función $f(n)$ preestablecida por el usuario.

Esta detección consiste de los siguientes dos pasos:

1. Construir un agrupamiento conceptual de los grafos,
2. Recorrer el agrupamiento conceptual y detectar, para cada grupo, las desviaciones correspondientes.

A continuación se describen las principales características del agrupamientos conceptual, posteriormente se explica su aplicación en la detección de las desviaciones.

3 Agrupamiento de los grafos conceptuales

En algunos trabajos previos se presento un método para el agrupamiento conceptual de un conjunto de grafos conceptuales (Montes-y-Gómez *et al.*, 2001b). En estos trabajos se argumentó que la jerarquía conceptual resultante indica la organización estructura oculta de la colección de grafos, pero que además constituye un resumen o índice de la colección, que facilita el descubrimiento de otros tipos de patrones interesantes, por ejemplo las desviaciones contextuales. A continuación se describe brevemente las principales características de la jerarquía conceptual

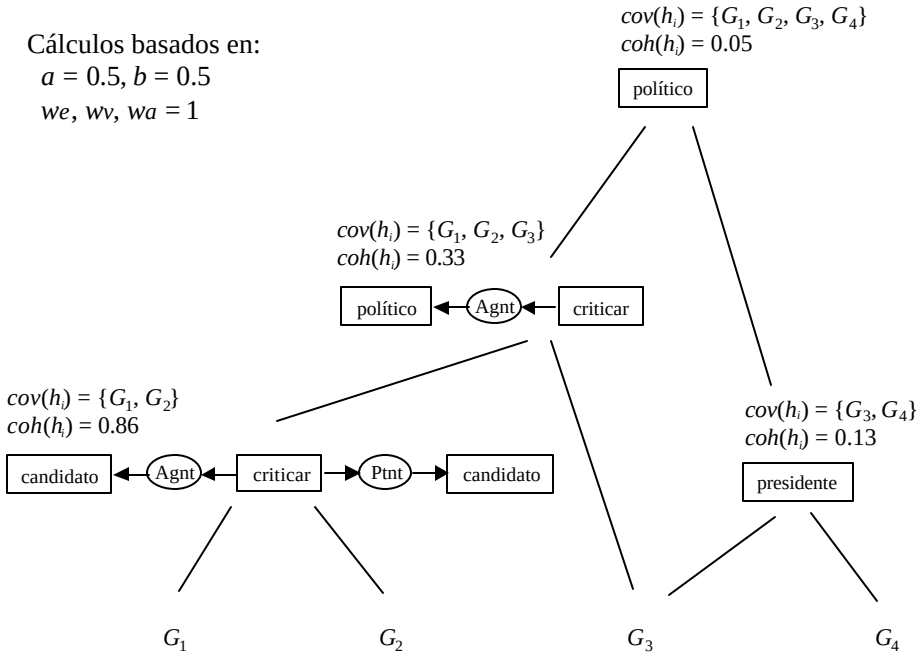

Figura 2. Agrupamiento conceptual del conjunto de grafos.

El *agrupamiento conceptual*, a diferencia de las técnicas tradicionales de agrupamiento, no solo permite dividir el conjunto de grafos conceptuales en varios grupos, sino también asociar una descripción a cada uno de estos grupos y organizarlos jerárquicamente de acuerdo con dichas descripciones (Michalski, 1980).

Básicamente, dado un conjunto de grafos conceptuales, este método identifica todas sus *regularidades* –elementos comunes a dos o más grafos– del conjunto, y además construye una jerarquía conceptual de ellas. Por ejemplo, dado el pequeño conjunto de grafos conceptuales de la figura 1, este método construye la jerarquía de la figura 2.

La jerarquía resultante no es necesariamente un árbol o *lattice*, sino un conjunto de árboles (esto es, un bosque). Esta jerarquía es una especie de red de herencia, en donde los nodos inferiores indican regularidades especializadas y los nodos superiores sugieren regularidades generalizadas.

Formalmente, cada nodo h_i de esta jerarquía se representa por una triada[2] ($cov(h_i)$, $desc(h_i)$, $coh(h_i)$), donde:

- $cov(h_i)$, llamada cobertura de h_i, es el conjunto de grafos cubiertos por la regularidad h_i.
- $desc(h_i)$, llamada descripción de h_i, es el conjunto de los elementos comunes de los grafos cubiertos por h_i, es decir, es el *traslape* de los grafos de $cov(h_i)$.

[2] Esta notación fue adaptada de (Bournaud and Ganascia, 1996), donde cada nodo se representa mediante un par ($cov(h_i)$, $desc(h_i)$).

- $coh(h_i)$, llamada cohesión de h_i, es la semejanza mínima entre dos grafos cualesquiera de $cov(h_i)$, esto es: $\forall G_i, G_j \in \text{cov}(h_i),\ sim(G_i, G_j) \geq coh(h_i)$.

Además, en esta jerarquía, el nodo h_j es un hijo del nodo h_i ($h_j \in S(h_i)$), si y solo si:

- El nodo h_i agrupa o cubre más grafos que el nodo h_j: $\text{cov}(h_j) \subset \text{cov}(h_i)$.
- La descripción del nodo h_i es una generalización de la descripción del nodo h_j: $desc(h_j) < desc(h_i)$.
- La cohesión de los grafos del agrupamiento h_i es menor o igual que la cohesión de los grafos del agrupamiento h_j: $coh(h_i) \leq coh(h_j)$.

4 Detección de las desviaciones contextuales

La detección de las desviaciones contextuales en un conjunto de grafos conceptuales $C = \{G_i\}$ se auxilia de su jerarquía conceptual H.

En esta jerarquía, cada nodo h_i indica un contexto específico del conjunto de grafos conceptuales C. Este contexto se describe por la regularidad $desc(h_i)$, y se compone por el conjunto de grafos $cov(h_i)$.

Además, el conjunto de nodos hijo de h_i indica una partición del contexto h_i, donde cada una de las descripciones de sus nodos $desc(h_j)$ expresa una característica, posiblemente representativa, del contexto h_i.

De acuerdo con esto y con las consideraciones de la sección 3, lo siguiente puede establecerse:

Característica representativa: La descripción $desc(h_j)$ del nodo $h_j \in S(h_i)$ es una característica representativa del contexto h_i si:

$$\left|\text{cov}(h_j)\right| \geq f\left(\left|\text{cov}(h_i)\right|\right)$$

Grafo conceptual raro: El grafo conceptual $G_i \in cov(h_i)$ es raro en el contexto h_i, si y solo si, no existe ninguna característica representativa $desc(h_j)$ del contexto h_i tal que: $G_i \in \text{cov}(h_j)$.

El conjunto de los grafos conceptuales raros del contexto h_i se denota como $R(h_i)$.

Desviación contextual: El grafo conceptual $desc(h_k)$, relacionado con el nodo $h_k < h_i$, es una desviación del contexto h_i, si y solo si: $\forall G_i \in \text{cov}(h_k) \Rightarrow G_i \in R(h_i)$.

En este caso, la siguiente desviación contextual puede establecerse:

$$desc(h_i):\ desc(h_j)\ \left(r = \frac{\left|cov(h_j)\right|}{\left|cov(h_i)\right|},\ s = \frac{\left|cov(h_i)\right|}{\left|C\right|} \right)$$

Esta definición permite encontrar *todas* las desviaciones contextuales (de acuerdo con una función $f(n)$ preestablecida por el usuario) en un conjunto dado de grafos conceptuales. Muchas de estas desviaciones contienen información redundante, esto es,

información implícita en otras desviaciones. Por ejemplo, si es raro que se hable de aves en un conjunto determinado de grafos conceptuales, entonces obviamente es aún más raro que se hable de aves de rapiña. Entonces, es necesario eliminar las desviaciones contextuales redundantes.

Desviación contextual redundante: La desviación contextual $g_i : g_k(a, b)$ es redundante si existe otra desviación contextual $g_i : g_j(g, b)$, tal que $g_k < g_j$.

A continuación se describe el algoritmo general para la detección de las desviaciones contextuales en un conjunto de grafos conceptuales. Este algoritmo recorre, en forma descendente, todos los nodos de la jerarquía conceptual. Para cada nodo h_i, que indica un contexto específico de la colección, identifica sus características representativas y también sus grafos raros.

Después, a partir del conjunto de grafos raros, detecta los nodos descendientes de h_i ($h_j < h_i$), cuya descripción $desc(h_j)$ representa una desviación para el contexto h_i.

```
1     procedure DetectaTodasDesviaciones (H)
2           for each nodo h_i ∈ H: |cov(h_i)| == 1
3                 DesviacionesContextules (h_i)

1     procedure DesviacionesContextuales (h_c)
2           if |cov(h_c)|/|C| ≥ minsup
3                 RARE ← cov(h_c)
4                 for each nodo h_s ∈ S(h_c)
5                       if |cov(h_s)| > log2 |cov(h_c)|
6                             RARE → cov(h_s)
7                 for each nodo h_s ∈ S(h_c)
8                       if |cov(h_s)| ≤  log2 |cov(h_c)|
9                             DesviacionMaxima (h_c, h_s, RARE)
10          for each nodo h_p ∈ P(h_c)
11                DesviacionesContextuales (h_p)

1     procedure DesviacionMaxima (h_c, h_d, RARE)
2           rareza  = |cov(h_d)| / |cov(h_c)|
3           soporte = |cov(h_c)| / |C|
4           if  h_d ⊆ RARE  &  rareza ≤ maxrar
5                 OUT ← "h_c: h_d  (rareza/soporte)"
6           else
7                 for each nodo h_s ∈ S(h_d)
8                       DesviacionMaxima (h_s)
```

Por ejemplo, consideremos el conjunto de grafos conceptuales de la figura 1 y su agrupamiento conceptual de la figura 2 con la función $f(n) = 0.4$:

- Con el contexto [político], las características representativas son [político]←(agt)←[criticar] y [presidente], no hay grafos raros y por lo tanto tampoco desviaciones.
- Con el contexto [político]←(agt)←[criticar], la característica representativa es [candidato]←(agt)←[criticar]→(ptn)→[candidato], en este caso G3 es raro y por lo tanto la desviaciones es [presidente:Castro]←(agt)←[criticar]→(ptn)→[elección].

- Con otros contextos, todos los hijos indican características representativas y no hay desviaciones.

Entonces, la única desviación contextual es:

[político]←(agt)←[criticar]:[presidente:Castro]←(agt)←[criticar]→(ptn)→[elección] (.33/.75)

Conclusiones

La detección de desviaciones es un problema importante de la minería de datos. En este articulo se introdujo el problema de descubrir desviaciones en un conjunto de datos semiestructurados; en particular, en un conjunto de grafos conceptuales.

Típicamente, la detección automática de desviaciones se realiza aplicando métodos de tipo estadístico o basados en distancia. A diferencia de ellos, el método aquí propuesto se basa en las *regularidades*. Es decir, este método considera que toda característica común a varios grafos de la colección es una característica representativa, y define como raro cualquier grafo que no comparte ninguna de estas características.

En general este método tiene las siguientes características novedosas:

1. Permite detectar tanto los grafos raros del conjunto dado, como los *patrones raros* en dicho conjunto. Estos últimos señalan las características comunes y exclusivas de los grafos raros, y por lo tanto son una manera *resumida* de expresar las desviaciones del conjunto de grafos.
2. *Aprovecha el agrupamiento conceptual* de los grafos para identificar y construir los patrones raros.
3. Considera las desviaciones con respecto a diferentes contextos (subconjuntos) del conjunto de grafos, lo que permite visualizar las desviaciones desde *diferentes perspectivas* y a *diferentes niveles de generalización*.

Agradecimientos

Este trabajo se realizó con el apoyo parcial del CONACyT, CGEPI-IPN, y SNI, México.

Referencias

1. Arning, Agrawal and Raghavan (1996), A Linear Method for Deviation Detection in Large Databases, Proc. of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, 1996.
2. Barnett and Lewis (1994), Outliers in Statistical Data, New York: John Wiley & Sons, 1994.
3. Bournaud and Ganascia (1996), Conceptual Clustering of Complex Objects: A Generalization Space based Approach, Lecture Notes in Artificial Intelligence 954, Springer, 1996.
4. Bournaud and Ganascia (1997), Accounting for Domain Knowledge in the Construction of a Generalization Space, Lectures Notes in AI (1257), Springer-Verlag, 1997.

5. Ellis and Lehmann (1994), Exploiting the Induced Order on Type-Labeled Graphs for fast Knowledge Retrieval, Conceptual Structures: Current Practices, William M. Tepfenhart, Judith P. Dick and John F. Sowa Eds., Lecture Notes in Artificial Intelligence 835, Springer-Verlag 1994.
6. Genest and Chein (1997), An Experiment in Document Retrieval using Conceptual Graphs, Conceptual structures: Fulfilling Peirce´s Dream. Lecture Notes in artificial Intelligence 1257, Springer 1997.
7. Godin, Mineau and Missaoui (1995), Incremental Structuring of Knowledge Bases, International KRUSE Symposium, August 11-13, Santa Cruz, California, 1995.
8. Guzmán (1996), Uso y Diseño de Mineros de Datos, Soluciones Avanzadas, Num. 34, 1996.
9. Han and Kamber (2001), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001.
10. Huibers, Ounis and Chevallet (1996), Conceptual Graph Aboutness, Conceptual Structures: Knowledge Representation as Interlingua. Peter W. Elkland, Gerard Ellis, Graham Mann Eds., Lecture Notes in Artificial Intelligence, Springer, 1996.
11. Knorr and Ng (1998), Algorithms for Mining Distance-based Outliers in Large Datasets, Proc. of the International Conference on Very Large Data Bases (VLDB'98), Newport Beach, CA, 1997.
12. Michalski (1980), Knowledge Acquisition thorough Conceptual Clustering: A Theoretical Framework and Algorithm for Partitioning Data into Conjunctive Concepts, International Journal of Policy Analysis and Information Systems, Vol. 4, 1980.
13. Mineau and Godin (1995), Automatic Structuring of Knowledge Bases by Conceptual Clustering, IEEE Transactions on Knowledge and Data Engineering, 7(5), 1995.
14. Montes-y-Gómez, López-López, Gelbukh (2000a). Information Retrieval with Conceptual Graph Matching. Proc. DEXA-2000, 11th International Conference and Workshop on Database and Expert Systems Applications, Greenwich, England,. Lecture Notes in Computer Science 1873, Springer-Verlag September 2000, pp. 312–321.
15. Montes-y-Gómez, Gelbukh and López-López (2000b), Comparison of Conceptual Graphs, Proc. MICAI-2000, 1st Mexican International Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence 1793, Springer 2000.
16. Montes-y-Gómez, Gelbukh, López-López and Baeza-Yates (2001a), Flexible Comparison of Conceptual Graphs, Proc. of DEXA-2001, 12th International Conference and Workshop on Database and Expert Systems Applications, Lecture Notes in Computer Science, Springer-Verlag, September 2001.
17. Montes y Gómez, Gelbukh, López López, and Baeza-Yates (2001b). Text mining with conceptual graphs. Proc. NLPKE-2001, Mini symposium on Natural Language Processing and Knowledge Engineering at SMC-2001, Systems, Man, And Cybernetics, IEEE, 2001.

18. Myaeng, (1992), Using Conceptual graphs for Information Retrieval: A Framework for Adequate Representation and Flexible Inferencing, Proc. of Symposium on Document Analysis and Information Retrieval, Las Vegas, March 1992.
19. Myaeng and López-López (1992), Conceptual Graph Matching: a Flexible Algorithm and Experiments, Journal of Experimental and Theoretical Artificial Intelligence, Vol. 4, 1992, pp. 107-126.
20. Mugnier (1995), On generalization/specialization for conceptual graphs, Journal of Experimental and Theoretical Artificial Intelligence, Vol. 7, 1995.
21. Mugnier and Chein (1992), Polynomial Algorithms for Projection and Matching, Proc. of thee 7th Conceptual Graphs Workshop, Las Cruces, NM, 1992.
22. Sowa (1984), Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, M.A., 1984.
23. Sowa (1999), Knowledge Representation: Logical, Philosophical and Computational Foundations, First Edition, Thompson Learning, 1999.

Detection of abnormal patterns in a semi-structured data set*

M. Montes-y-Gómez [1], **A. Gelbukh** [1], and **A. López-López** [2]

[1] Center for Computing Research (CIC-IPN), Mexico.
mmontesg@susu.inaoep.mx, gelbukh@cic.ipn.mx

[2] National Institute of Astrophysics, Optics, and Electronics (INAOE), Mexico.
allopez@inaoep.mx

Abstract. Detection of deviations is an important problem in data mining. We consider the discovery of the deviations in a group semi-structured data set; in particular, in data represented by conceptual graphs. Typically, the automatic detection of deviations is done with statistical or distance-based methods. Our method, however, is based on the regularities, i.e., on characteristics common to several graphs in the collection. The method discovers the graphs that do not share any of these regularities. Also, the abnormal patterns are detected in the set, which is a way to summarize the deviations found. Such patterns are discovered using the conceptual clustering of the graphs. The method allows considering the deviations regarding different contexts (subsets) at the user's choice, i.e., from different perspectives and at different generalization levels.

Keywords: data mining, text mining, conceptual graphs, deviations, clustering.

* Work done with partial support of CONACyT, CGEPI-IPN, and SNI, Mexico.