M. Montes-y-Gómez, A. Gelbukh, A. López-López. *Detecting deviations in text collections: An approach using conceptual graphs*. Proc. MICAI-2002: Mexican International Conference on Artificial Intelligence. Lecture Notes in Artificial Intelligence N 2313, Springer-Verlag, 2002.

Detecting Deviations in Text Collections: An Approach using Conceptual Graphs

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

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

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

Abstract. Deviation detection is an important problem of both data and text mining. In this paper we consider the detection of deviations in a set of texts represented as conceptual graphs. In contrast with statistical and distance-based approaches, the method we propose is based on the concept of generalization and regularity. Among its main characteristics are the detection of rare patterns (that attempt to give a generalized description of rare texts) and the ability to discover local deviations (deviations at different contexts and generalization levels). The method is illustrated with the analysis of a set of computer science papers.

Keywords: natural language processing, deviation detection, text mining, conceptual graphs, regularity.

1 Introduction

For our civilization, knowledge is the most valuable treasure. Most of this knowledge exists in the form of natural language as books, journals, reports, etc. Therefore, the real possession of all this knowledge depends on our capabilities for doing different tasks with texts, for instance: search for interesting texts, compare different texts, or summarize them.

Text mining, the new research area of text processing, is focused in this kind of tasks. Mainly, it is concerned with the discovery of interesting patterns, such as clusters, associations, *deviations*, similarities and differences from text collections (Feldman, 1999; Mladenic, 2000; Ciravegna *et al.*, 2001).

Currently, most methods of text mining use simple and shallow representations of the texts. On one hand, such representations are easily extracted from texts and easily analyzed, but on the other hand, they usually restrict the discovered patterns to the topic level.

To make text mining more useful, richer representations than just keywords, i.e., representations with more types of textual elements, must be used. On the basis of this idea, we are designing a method for doing text mining at detail level (Montes-y-Gómez, 2002). This method uses *conceptual graphs* (Sowa, 1999) for representing the texts content, and considers different tasks, such as: clustering generation, association discovery and deviation detection.

This paper focuses on the detection of rare patterns – also called deviations – among a set of conceptual graphs representing document details. Basically, it presents a method that detects deviations based on the concept of *regularity* – instead of distance – and that allows detecting local deviations and deviations at different levels of generalization.

The paper is organized as follows. Section 2 describes some previous work on deviation detection. Section 3 introduces conceptual graphs and their clustering. Section 4 presents a method for detecting deviations in a set of conceptual graphs. Then Section 5 shows some experimental results from the analysis of a set of computer science papers. Finally, section 6 draws some preliminary conclusions.

2 Related Work

Traditional statistical methods consider deviations as a source of noise, and try to reduce their effects. On the contrary, recent data mining methods consider the deviations a especially interesting hidden knowledge about data. These methods are mainly of two kinds: methods that use additional information about the data (e.g. Guzmán, 1996), and methods that take advantage of the data's own redundancy (Han and Kamber, 2001). Among the latter we distinguish the following two approaches:

- *The statistical approach* assumes a distribution model for the data, and then identifies deviations with respect to the model using a discordancy test. The application of this test requires knowledge about the data (such as the data distribution) and knowledge about the distribution parameters (such as mean and variance). This approach is presented in (Barnett and Lewis, 1994).
- *The distance-based approach* considers that an object *o* in a data set *S* is a deviation with parameters *p* and *d*, if at least a fraction *p* of the objets in *S* lie at a distance greater than *d* from *o*. The application of this approach requires a distance measure between the objects of the problem and some parameters, such as the number of neighbors of a given object in order to be not considered as a deviation (Knorr and Ng, 1998).

In text mining two different approaches are used for detecting deviations. The first approach focuses on detecting rare texts in a given collection (Alexandrov *et al.*, 2000). The second allows detecting rare topics in a collection (Feldman and Dagan, 1995; Allan *et al.*, 1998).

3 Background

3.1 Conceptual graphs

A conceptual graph is a bipartite graph (Sowa, 1999) with two different kinds of nodes: concepts and relations.

- *Concepts* represent entities, actions, and attributes. Concept nodes have two attributes: type and referent. Type indicates the class of the element represented by the concept. Referent indicates the specific instance of the class referred to by the node. Referents may be generic or individual.
- *Relations* show the inter-relationships among the concept nodes. Relation nodes also have two attributes: valence and type. Valence indicates the number of concepts involved in the relation, while the type expresses its semantic role.

For instance, the graph

[cat:Tom]←(agt)←[chase]→(ptn)→[mouse]→(atr)→[brown]

represents the phrase "*Tom is chasing a brown mouse*". It has three concepts and three relations. The concept [cat: Tom] is an individual concept of the type *cat* (a specific cat Tom), while the concepts [chase] and [mouse] are generic concepts. All relations in this graph are binary. For instance, the relation (attr) for *attribute* indicates that the mouse has brown color. The other two relations stand for *agent* and *patient* of the action [chase].

3.2 Conceptual clustering

In some previous work, we presented a method for *conceptual clustering* of conceptual graphs (Montes-y-Gómez *et al.*, 2001). There, we argued that the resulting conceptual hierarchy expresses the hidden organization (structure) of the collection of graphs, but also constitutes an abstract or *index* of the collection that facilitate the discovery of other hidden patterns, e.g. the contextual deviations. Following, we briefly explain the main characteristics about this conceptual hierarchy.

Conceptual clustering –unlike the traditional cluster analysis techniques– allows not only to divide the set of graphs into several groups, but also to associate a description to each group and to organize them into a hierarchy. The resulting hierarchy H is not necessarily a tree or lattice, but a set of trees (a forest). This hierarchy is a kind of inheritance network, where those nodes close to the bottom indicate specialized regularities and those close to the top suggest generalized regularities[1]. In section 5, we show part of a cluster hierarchy built from a set of computer science papers.

Formally, each node $h_i \in H$ is represented by a triplet ($cov(h_i)$, $desc(h_i)$, $coh(h_i)$). Here $cov(h_i)$, the coverage of h_i, is the set of graphs covered by the regularity h_i; $desc(h_i)$, the description of h_i, consists of the common elements of the graphs of $cov(h_i)$; $coh(h_i)$, the cohesion of h_i, indicates the least similarity among any two graphs of $cov(h_i)$.

Also, the node h_i is an antecessor of the node h_j if: $cov(h_j) \subset cov(h_i)$, $desc(h_j) < desc(h_i)$ and $coh(h_j) \geq coh(h_i)$.

4 Conceptual graph deviations

4.1 Basic considerations

Our method, different to the statistical and distance-based approaches, is based on the concept of *regularity*. Basically, it defines any object (graph or text in our case) without a representative characteristic –a characteristic that is common to a great number of its "neighbors"– as a rare object, and consequently as a possible deviation.

[1] The construction of the conceptual hierarchy is a knowledge-based procedure (Montes-y-Gómez *et al.*, 2001). Basically, a concept hierarchy (defined by the user in accordance with his interests) handles the generalization/specialization of the graphs when the conceptual hierarchy is constructed.

This approach is similar to that proposed by Arning *et al*. (1996), who define a deviation as those elements that increase the variance of the complete set of objets.

The detection of deviations in a set of conceptual graphs is supported on the following ideas. Given a set of conceptual graphs, $C = \{G_i\}$:

- A *representative characteristic* is any common generalization g_c of more than m conceptual graphs of the set, where m is a given threshold. Let denote by F the set of representative characteristics of C.
- A *rare conceptual graph* is a graph that has no representative characteristic[2]. Thus, the set of rare graphs is defined as: $R = \{G_r \in C | \nexists g_c \in F : G_r < g_c\}$.
- A deviation d is a pattern that describes one or more of the rare graphs. In other words, a deviation is a generalization of some rare graphs of C. Thus, given a deviation d the following conditions are satisfied:

  1. $\exists G_r \in R : G_r < d$.
  2. $\nexists G \in C, \nexists g \in F : G < g \wedge G < d$.

Therefore, given a set of conceptual graphs C, a *contextual deviation* is an expression of the form: $g_i : g_j \, (r, s)$. In this expression g_i is the context and g_j is the description of the rare graphs (a m-deviation); r is the rarity of the deviation on the context, and s is the support of the context with respect to the whole set. For instance, the following deviation indicates that just 4% of the graphs about animals (of some imaginary set) mention a bird of prey, while 32% of the graphs of the entire set are about animals: [animal]: [bird]→ (kind)→[prey] (4%,32%).

4.2 Method for detecting the deviations

Detecting deviations in a given set of graphs is defined as the problem of finding all $g_i : g_j \, (r, s)$ expressions according with the user defined threshold m.

The process of detection of the deviations in a set of conceptual graphs C is based on the existance of a conceptual clustering H. Each node h_i of this hierarchy represents a different context (related subset) of C, groups the graphs of $cov(h_i)$ and is described by the conceptual graph $desc(h_i)$.

For each context (node of the hierarchy), we detect deviations based on the following definitions:

Representative characteristic: The description $desc(h_j)$ of the node $h_j < h_i$ is a *representative characteristic* for the context h_i if $|cov(h_j)| \geq m \times |cov(h_i)|$.

Rare conceptual graph: The graph $G_i \in cov(h_i)$ is a *rare graph* for the context h_i if there is no representative characteristic $desc(h_j)$ of context h_i such that $G_i \in cov(h_j)$. The set of rare conceptual graphs of the context h_i is denoted by $R(h_i)$.

[2] If there is no representative characteristic, then it is not possible to detect any deviation.

```
***Procedure Detect_Deviations in H***

*Parameters m*

for each node $h_i$ of the hierarchy $H$
  set NOT_RARE ← ∅
  for each son-node $h_s$ of $h_i$
    if $|cov(h_s)| \geq m$ X $|cov(h_i)|$
       Insert in NOT_RARE the graphs covered by $h_s$
  for each son-node $h_s$ of $h_i$
    if $|cov(h_s)| < m$ X $|cov(h_i)|$
      if the node $h_s$ does not cover any graph of NO_RARE
        Define rarity $r \leftarrow |cov(h_s)|/|cov(h_i)|$
        Define support $s \leftarrow |\mathrm{cov}(h_i)|/|C|$
        Build deviation $h_i$ : $h_s$  $(r, s)$.
      else
        Maximal_Deviations of $h_i$ considering $h_s$

***Procedure Maximal_Deviation of h_i considering h_k***

*Parameters $h_i$, $h_j$,* NOT_RARE *of $h_i$*

for each son-nodo $h_s$ of $h_k$
  if $h_k$ does not cover any graph of NO_RARE
    Define rarity $r \leftarrow |cov(h_k)|/|cov(h_i)|$
    Define support $s \leftarrow |cov(h_i)|/|C|$
    Build deviation $h_i$ : $h_k$  $(r, s)$.
  else
    Maximal_Deviation of $h_i$ considering $h_k$
```

Figure 1. Deviation detection algorithm.

Contextual deviation: The description *desc*(h_k) of the node $h_k < h_i$ is a *contextual deviation* of the context h_i if $\forall G_i \in cov(h_k)$ it holds $G_i \in R(h_i)$. In this case, the contextual deviation is expressed as:

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

These definitions allow finding *all* contextual deviations for a given set of graphs, and also for a given *m*. Many of these deviations are redundant. For instance, if the graphs that mention birds are rare then the graphs about birds of prey are also rare. Thus, it is necessary to eliminate the redundant deviations.

Redundant deviation: The contextual deviation $g_i : g_k\,(a, b)$ is *redundant* if there is another contextual deviation $g_i : g_j\,(g, b)$ such that $g_k < g_j$.

The basic algorithm for detecting deviations in a set of conceptual graphs – in their conceptual hierarchy – works as follows. It traverses all the hierarchy (using a *bottom-up* approach) and for each node h_i (a context) defines the set of representative

Logical Analysis of Programs

The first part of the paper is devoted to techniques for the automatic generation of invariants. The second part provides criteria for using the invariants to check simultaneously for correctness (including termination) or incorrectness. A third part examines the implications of the approach for the automatic diagnosis and correction of logical errors.

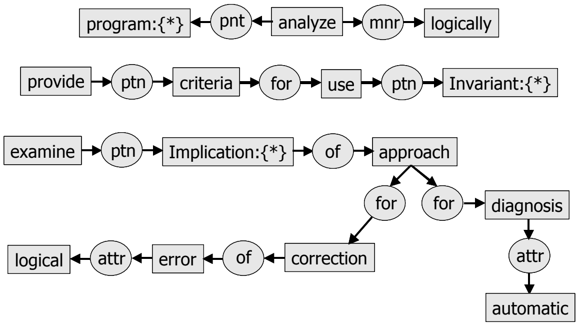


Figure 2. A scientific paper and it conceptual graph

characteristics and rare elements. Then, it detects the nodes $h_j < h_i$ whose descriptions $desc(h_j)$ represent a contextual deviation for the given context h_i. The algorithm is described in figure 1.

5 Experimental results

This section describes the analysis of a set of conceptual graphs that represents the content of 495 paper surrogates of computer science. Figure 2 shows a paper surrogate and its corresponding conceptual graph. The method to extract and build the graphs is described in (Tapia-Melchor and López-López, 1998; Montes-y-Gómez, 1999).

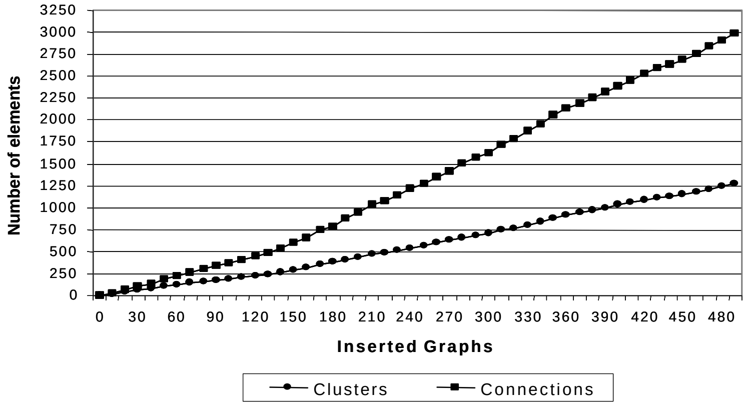
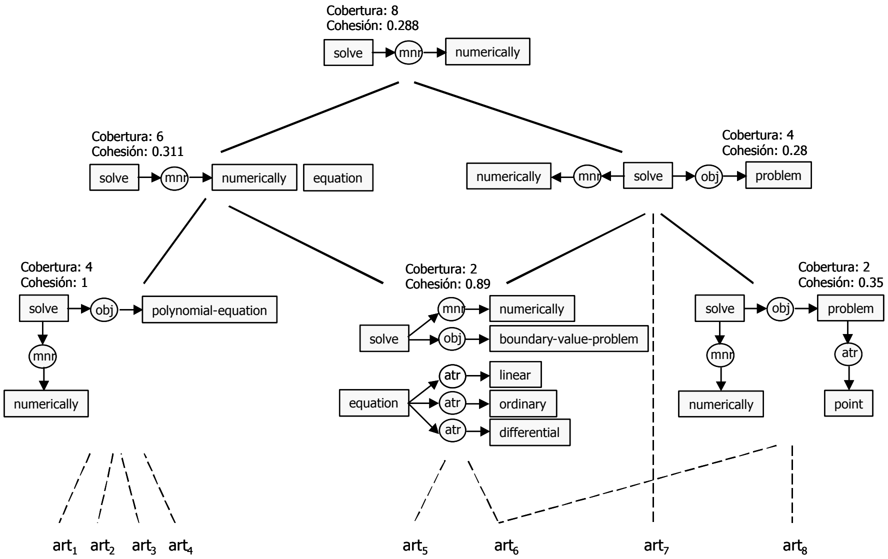


Figure 3. Growth of the cluster hierarchy



art_1 a art_4 ...(numerical solution of the polynomial equation)...
art_5 ...(the numerical solution of boundary value problems for linear ordinary differential equations)...
art_6 ...(the numerical solution of an n-point boundary value problem for linear ordinary differential equations)...
art_7 - ...(the numerical solution of a thin plate heat transfer problem)...
art_8 - ...(the numerical solution of nonlinear two-point boundary problems by finite difference methods)...

Figure 4. A part of the cluster hierarchy

In order to detect the deviations in the given set of graphs, we first built it conceptual clustering. The experiments demonstrate that this kind of clustering, when the graphs represent text details, is practical. For instance, figure 3 shows an almost linear growth in the number of clusters and connections of the conceptual hierarchy. Also, this clustering is rich enough for discovering patterns, since it maintains most conceptual and relational information. The figure 4 shows a small part of the resulting hierarchy.

For the experiments we use $0.1 \leq m \leq 0.5$. The best case was obtained with $m = 0.25$. We detected 23 deviations; two examples are showed in Figure 5. They indicate that the papers are focused on the description of different *procedures*, being the *data division* the procedure less studied. Also, some papers consider the *solution of equations*, but just very few study the solution of *polynomial equations by the Barstow-Hitchcock method*.

6 Conclusions

Current methods of text mining use simple and shallow representation of texts, for instance, a list of keywords. On one hand, such representations are easily extracted from texts and easily analyzed, but on the other hand, they usually restrict the expressiveness and diversity of the discoveries.

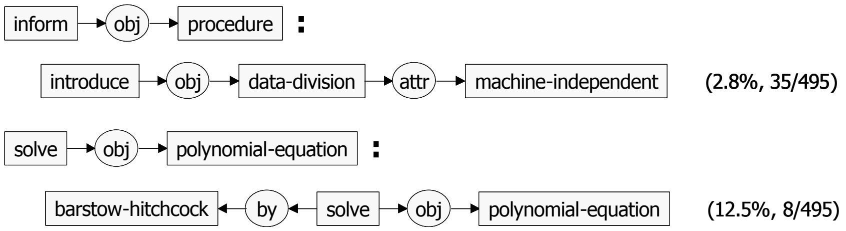


Figure 5. Two deviations of the set of computer science papers.

Our research is focused on this problem. Basically, we proposed to use conceptual graphs for representing the text content, and developed some methods to analyze this kind of representations and discover detail patterns among texts.

In this paper we present a method for detecting rare patterns –deviations– in a given set of conceptual graphs. This method is different from traditional statistical and distance-based approaches, because it detects the deviations based on the concept of regularity. Some important characteristics of this method are:

- Detects not only rare graphs but also patterns about them. These patterns summarize the content of the rare graphs.
- Uses the conceptual clustering of the graphs as an index of the collection. This strategy facilitates the detection of local deviations.
- Identifies deviations for different contexts of the set of graphs (local deviations), and thus allows visualizing the deviations from different perspectives and at different levels of generalization.

Acknowledgements. This work was done under partial support of CONACyT, CGEPI-IPN, and SNI, Mexico.

References

1. Alexandrov, M., A. Gelbukh, and P. Makagonov (2000), On Metrics for Keyword-Based Document Selection and Classification, Proc. of the Conference on Intelligent Text Processing and Computational Linguistics CICLing-2000, Mexico City, Mexico, February 2000.
2. Allan, Papka and Lavrenko (1998), On-line new Event Detection and Tracking, Proc. of the 21st ACM-SIGIR International Conference on Research and Developement in Information Retrieval, August 1998.
3. 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.
4. Barnett and Lewis (1994), Outliers in Statistical Data, New York: John Wiley & Sons, 1994.
5. Ciravegna et al., Ed. (2001), Proc. of the 17Th International Joint Conference on Artificial Intelligence (IJCAI-2001), Workshop of Adaptive Text Mining, Seattle, WA, 2001.

6. Feldman and Dagan (1995), Knowledge Discovery in Textual databases (KDT), Proc. of the 1st International Conference on Knowledge discovery (KDD_95), pp.112-117, Montreal, 1995.
7. Feldman, Ed. (1999), Proc. of The 16th International Joint Conference on Artificial Intelligence (IJCAI-1999), Workshop on Text Mining: Foundations, Techniques and Applications, Stockholm, Sweden, 1999.
8. Guzmán (1996), Uso y Diseño de Mineros de Datos, J. Soluciones Avanzadas, Num. 34, 1996.
9. Han and Kamber (2001), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001.
10. 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.
11. Mladenic, Ed. (2000), Proc. of the Sixth International Conference on Knowledge Discovery and Data Mining, Workshop on Text Mining, Boston, MA, 2000.
12. Montes-y-Gómez, Gelbukh, López-López (1999), Document intentions expressed in titles. Extraction, representation, and possible use, Selected Works 1997-1998, Center for Computing Research (CIC-IPN), 1999.
13. Montes-y-Gómez, Gelbukh, López-López, Baeza-Yates (2001), Un Método de Agrupamiento de Grafos Conceptuales para Minería de Texto, J. Procesamiento de Lenguaje Natural, Vol. 27, Septiembre 2001.
14. Montes-y-Gómez (2002), Minería de texto usando la semejanza entre estructuras semánticas, Ph.D. thesis, Center for Computing Research (CIC-IPN), Mexico, 2002.
15. Tapia-Melchor and López-López (1998), Automatic Information Extraction from Documents in WWW, Séptimo Congreso Internacional de Electrónica, Comunicaciones y Computadoras, CONIELECOMP 98, Febrero, 1998.
16. Sowa (1999), Knowledge Representation: Logical, Philosophical and Computational Foundations, 1st edition, Thomson Learning, 1999.