TEXT MINING WITH CONCEPTUAL GRAPHS

M. MONTES-Y-GÓMEZ [1], A. GELBUKH [1], A. LÓPEZ-LÓPEZ [2], and **R. BAEZA-YATES [3]**

[1] Centro de Investigación en Computación (CIC-IPN), Mexico.
mmontesg@susu.inaoep.mx, gelbukh@cic.ipn.mx
[2] Instituto Nacional de Astrofísica, Optica y Electrónica (INAOE), Mexico.
allopez@inaoep.mx
[3] Departamento de Ciencias de la Computación, Universidad de Chile, Chile.
rbaeza@dcc.uchile.cl

Abstract

A method for conceptual clustering of a collection of texts represented with conceptual graphs is presented. It uses the incremental strategy to construct the cluster hierarchy and incorporates some characteristics attractive for text mining proposes. For instance, it considers the structural information of the graphs, uses domain knowledge to detect the clusters with generalized descriptions, and uses a user-defined similarity measure between the graphs.

Keywords

Conceptual clustering, similarity measure, user interests profile, domain knowledge, summarization.

1 Introduction

Text mining has emerged as a new research area of text processing. It is focused on the discovery of new facts and world knowledge from large collections of texts that – unlike the problem of natural language understanding – do not explicitly contain the knowledge to be discovered (Hearst, 1999). The goals of text mining are similar to those of data mining: for instance, it also attempts to find clusters, uncover trends, discover associations and detect deviations in a large set of texts. Text mining has also adopted techniques and methods of data mining, e.g., statistical techniques and machine learning approaches.

The general framework of text mining consists of two main phases: a pre-processing phase and a discovery phase (Tan, 1999). In the first phase, the free-form texts are transformed to some kind of semi-structured representation that allows their automatic analysis, and in the second one, the intermediate representations are analyzed and some interesting and non-trivial patterns are hopefully discovered.

Many of the current methods of text mining use simple and shallow representations of texts. On one hand, such representations are easily extracted from the texts and easily analyzed, but on the other hand, they restrict the kind of discovered knowledge.

Recently in all text-oriented applications, including text mining, there is a tendency to start using more complete representations than just keywords, i.e. representations with more types of textual elements. In text mining, for instance, there is the belief that these new representations will expand the kinds of discovered knowledge (Hearst, 1999; Tan, 1999).

In this paper, we propose a text mining method that uses conceptual graphs (Sowa, 1984) as the intermediate representation of texts[1]. To achieve this goal, two problems are central: (1) the transformation of texts into conceptual graphs and (2) the automatic analysis of a collection of these graphs. This paper is not concerned with the first problem – the construction of the conceptual graphs is studied elsewhere (Sowa and Way, 1986; Baud *et al.*, 1992; Myaeng and Khoo, 1994; Montes-y-Gómez *et al.*, 1999); rather, it concentrates on the ***conceptual clustering*** of a set of conceptual graphs, and thus in the discovery of all their regularities.[2]

There are two well-known methods for the conceptual clustering of conceptual graphs (Mineau and Godin, 1995; Bournaud and Ganascia, 1996). The method described here, just like these two methods, follows an unsupervised learning strategy that incrementally builds a cluster hierarchy from the conceptual graphs. Additionally, this method incorporates some features that make it attractive for the text mining purposes. For instance, (1) it considers all structural information of the conceptual graphs; (2) it builds the cluster hierarchy emphasizing the interests

* Work partially supported by CONACyT, SNI, and CGEPI-IPN, Mexico.

[1] In this paper, a text may refer to a sentence, a paragraph or a complete document.

[2] A regularity is any common subgraph or generalization of two or more graphs of the set of graphs.

M. Montes y Gómez, A. Gelbukh, A. López López, Ricardo Baeza-Yates. *Text mining with conceptual graphs.* IEEE International Workshop on Natural Language Processing and Knowledge Engineering, NLPKE 2001, IEEE, pp. 898–903; www.Gelbukh.com/CV.

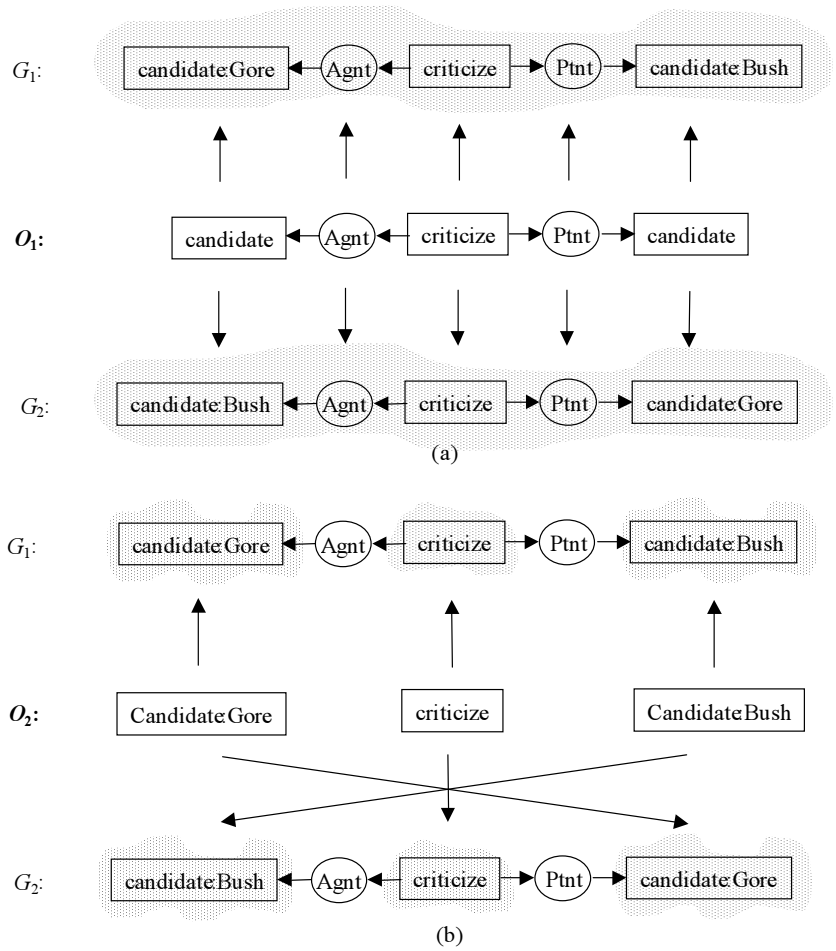

Figure 1. Different overlaps of two conceptual graphs

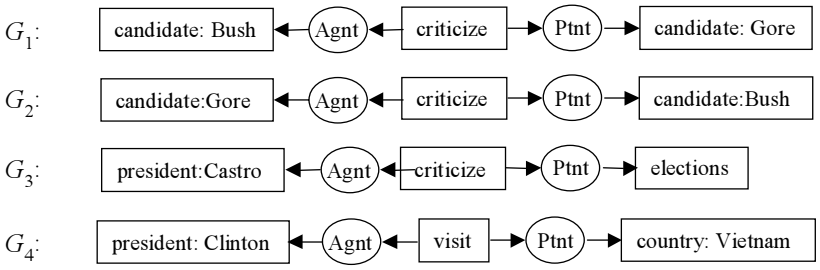

Figure 2. A small set of conceptual graphs

of the end-users; and (3) it uses domain knowledge, e.g. a thesaurus and some *is-a* hierarchies.

The rest of the paper is organized as follows. Section 2 briefly describes our approach for the comparison of two conceptual graphs. Section 3 defines the cluster hierarchy and also presents the incremental procedure for its construction. Finally, section 4 discusses some conclusions and future work.

2 Comparison of Conceptual Graphs

In order to cluster conceptual graphs, a means for comparing them is required. We have proposed a knowledge-based-user-oriented method for the comparison of two conceptual graphs (Montes-y-Gómez *et al.*, 2000; Montes-y-Gómez *et al.*, 2001). This method uses a thesaurus and some *is-a* hierarchies in

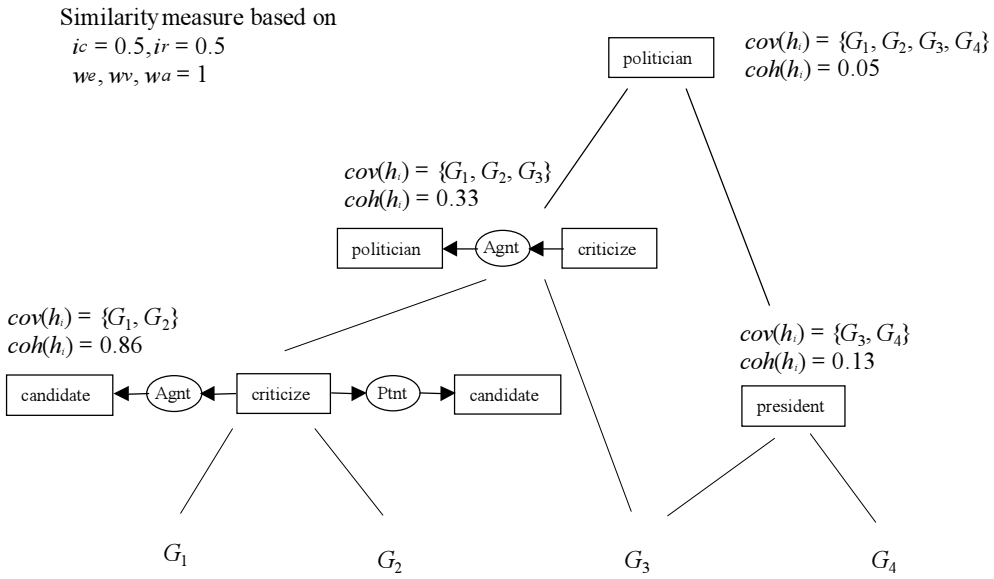

Figure 3. Conceptual clustering of the set of graphs

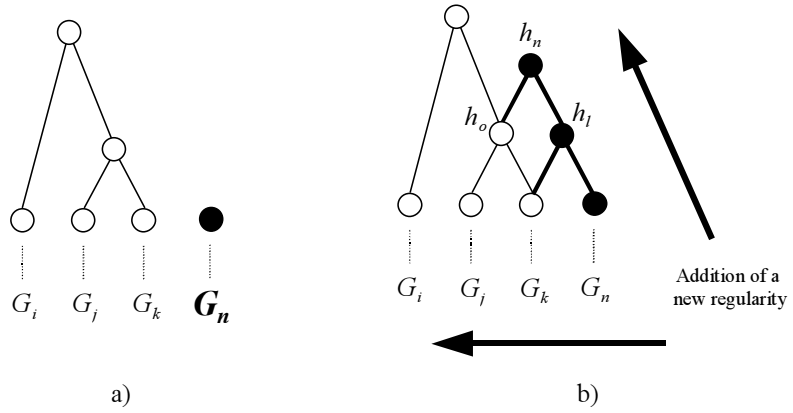

Figure 4. Incorporation of a new graph to the hierarchy

the construction of the cluster hierarchy. The use of this kind of information allows the identification of non-exact and generalized similarities.

Basically, this method builds a description of the similarity of the two graphs, and also obtains a measure of it. The description of the similarity is just an overlap of the graphs, i.e., a set of all their common but compatible generalizations.[3]

The similarity measure indicates the relative importance of one overlap with respect to the original graphs. It is defined as a combination of two values: the conceptual similarity and the relational similarity. The conceptual similarity depends on the common concepts of the two graphs, i.e, on how similar the entities, actions, and attributes mentioned in both conceptual graphs are. While the relational similarity expresses how similar the relations among the common concepts in the two conceptual graphs are, i.e., it indicates how similar the neighbors of the overlap in both original graphs are.

The similarity measure is computed in accordance to the user interests and is defined as follows:

$$sim(G_i, G_j) = f(w_E, w_V, w_A, i_c, i_r).$$

Here w_E, w_V, w_A are the importance of the entities, actions and attributes respectively, i_c indicates the importance of the conceptual similarity and i_r the importance of the structural similarity. The value of these parameters is restricted by the conditions: $w_E, w_V, w_A > 0$ and $i_c + i_r = 1$.

[3] In the cases, where there is more than one possible overlap between the graphs, that producing the greatest similarity measure is selected.

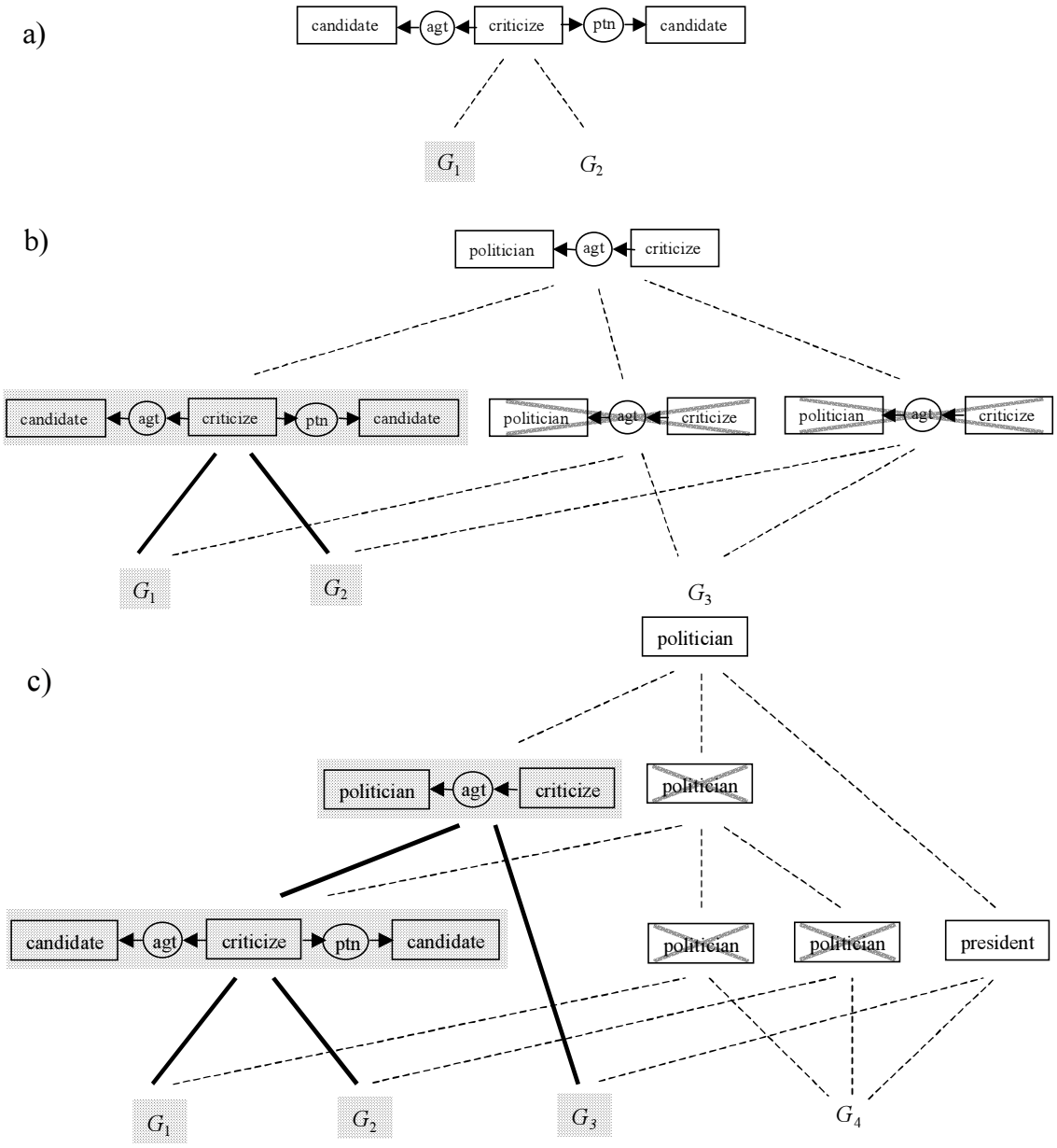

Figure 5. Hierarchy construction

Adjusting all these parameters, the method of comparison may be easily adapted to the different user specific discovery purposes. For instance, Figure 1 illustrates the comparison of two simple conceptual graphs; one indicating that Gore criticizes Bush, and the other that Bush criticizes Gore. These graphs have two possible overlaps O_1 and O_2. When the structural information is emphasized ($i_r > i_c$), then O_1 is considered the best description of the similarity (it expresses that a candidate criticizes another candidate), while when the conceptual information is stressed ($i_c > i_r$) then O_2 is a better description of it. This overlap expresses that both graphs mention Bush, Gore and the action of criticize.

3 Clustering of Conceptual Graphs

In order to find all the regularities of a given set of conceptual graphs, we use a conceptual clustering method. This method – unlike the traditional cluster analysis techniques – allows not only dividing the set of graphs into several groups, but also associating a description to each group and organizing them into a hierarchy. The resulting hierarchy is not necessarily a tree or lattice, but a set of trees (a forest). This hierar-

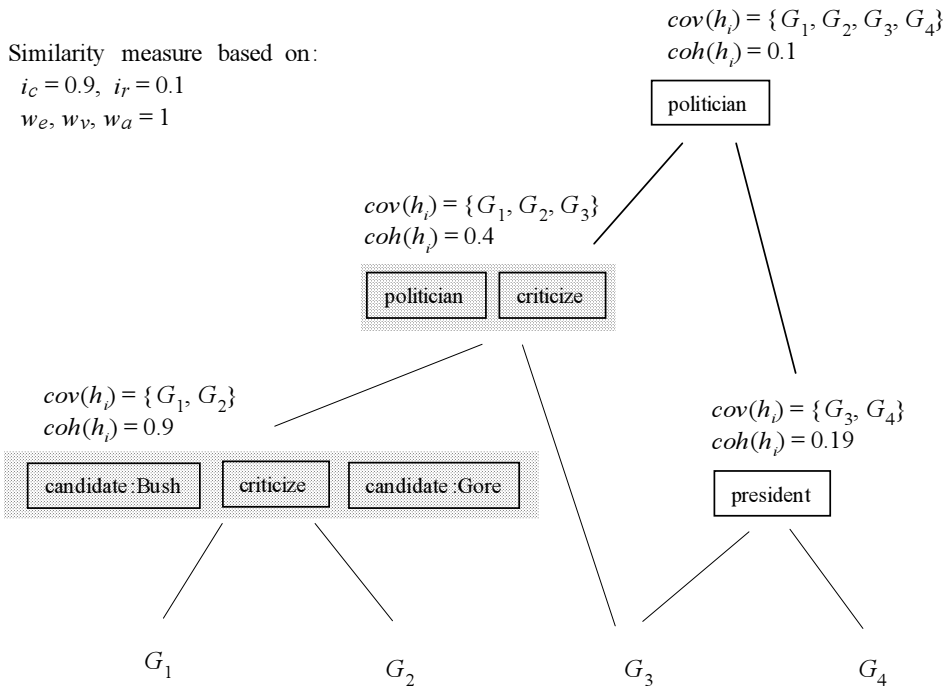

Figure 6. A different cluster hierarchy

chy is a kind of inheritance network, where the lower nodes indicate specialized regularities and the upper ones suggest generalized regularities. For instance, the hierarchy of Figure 3 expresses the conceptual clustering of the small set of graphs of Figure 2.

Formally, each node h_i of the cluster hierarchy is represented by a triplet[4] ($cov(h_i)$, $desc(h_i)$, $coh(h_i)$). 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 $cov(h_i)$, i.e., $desc(h_i)$ is the overlap of the graphs covered by h_i; $coh(h_i)$, the cohesion of h_i, indicates the less similarity among any two graphs of $cov(h_i)$, that is:

$$\forall G_i, G_j \in cov(h_i),\ sim(G_i, G_j) \geq coh(h_i)$$

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

3.1 Construction of the cluster hierarchy

Given a set of conceptual graphs, the construction of their cluster hierarchy is based on an incremental method. This method considers all the structural information of the conceptual graphs, and also uses the similarity measure among the graphs in order to emphasize the user interests.

In general, the incorporation of a new conceptual graph G_i into the hierarchy is done in two steps (see Figure 4). In the first step, a node covering only the new graph ($\{G_i\}$, G_i, 1) is added. In the second step, all the regularities related to the new graph are identified. These new regularities are added to the hierarchy in a *bottom-up* manner, i.e., each top-level node is formed by the combination of two nodes of a lower level. For instance, the node h_n of Figure 4 (b) was constructed from the nodes h_o and h_l. In this case, the node h_n was defined as follows:

$$\begin{aligned} \text{cov}(h_n) &= \text{cov}(h_o) \cup \text{cov}(h_l) \\ desc(h_n) &= match(desc(h_o), desc(h_l)) \\ coh(h_j) &= \begin{cases} sim(desc(h_o), desc(h_l)) & \text{if } |\text{cov}(h_o)| = |\text{cov}(h_l)| = 1 \\ \min(coh(h_o), coh(h_l)) & \text{otherwise} \end{cases} \end{aligned}$$

Here, the function $match(G_i,G_j)$ returns the best overlap – for the user purposes – of the graphs G_i and G_j, and $sim(G_i,G_j)$ computes their related similarity measure.

Also, when a new regularity is added to the hierarchy, all the duplicated regularities are deleted. For instance, if $desc(h_o) = desc(h_n)$, then the node h_o is deleted; while if $desc(h_l) = desc(h_n)$, then the node h_l is eliminated.

The process of construction of the hierarchy in Figure 3 is illustrated in the Figure 5. In this figure, the

[4] This notation was adapted from (Bournaud and Ganascia, 1996); where each node h_i was represented by a pair ($cov(h_i)$, $desc(h_i)$).

highlighted nodes correspond to the last-iteration hierarchy, and the rest of the elements indicate the comparisons done at the current iteration.

This example mainly shows the direct dependence existing between the construction of the cluster hierarchy and the comparison of conceptual graphs. This relation defines the conceptual clustering as a knowledge-based-user-oriented process. This means that the use of different knowledge bases (e.g. *is-a* hierarchies) and the setting of distinct user purposes (as different similarity measure parameters) may produce different cluster hierarchies.

For instance, the cluster hierarchy of Figure 6 is obtained when the conceptual similarities are stressed instated of the structural ones. In this figure, the highlighted nodes indicate the differences with the initial hierarchy (see Figure 3).

4 Conclusions and Future Work

We have presented a method for conceptual clustering of a collection of texts represented as a set of conceptual graphs. This method follows a traditional incremental strategy for the construction of the cluster hierarchy, but also incorporates some attractive characteristics for text mining proposes. For instance:

1. Considers all *structural information* of the graphs. This is a consequence of the method for the comparison of conceptual graphs (refer to [6,7]).
2. Uses *domain knowledge*. This basically allows the detection of clusters with generalized descriptions.
3. Uses the *similarity measure* of the graphs in order to emphasize the *user interests* during the hierarchy construction.

These characteristics allow incrementing the expression power of the clustering and the flexibility of it construction. Also, they define the conceptual clustering of the graphs as a knowledge-based and user-oriented process.

The resulting cluster hierarchy is an interesting discovery for two main reasons. On one hand, it indicates the implicit structure of the set of graphs (texts). On the other hand, it constitutes a kind of structured abstract of the set of graphs, which seems to be useful for the discovery of other kind of hidden patterns.

Some tasks for future work are: (1) the identification of the most important nodes of the hierarchy, i.e., the most interesting regularities in accordance with the user interests; (2) the discovery of association rules and contextual deviations using the cluster hierarchy as a special kind of index of the collection.

References

1. Baud, Rassinoux and Scherrer (1992), Natural Language Processing and Semantical Representation of Medical Texts, Meth Inform Med 31:117-25, 1992.
2. Bournaud and Ganascia (1996), Conceptual Clustering of Complex Objects: A Generalization Space based Approach, Lecture Notes in Artificial Intelligence 954, Springer, 1996.
3. Hearst (1999), Untangling Text Data Mining, To appears in Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, 1999.
4. Mineau and Godin (1995), Automatic Structuring of Knowledge Bases by Conceptual Clustering, IEEE Transactions on Knowledge and Data Engineering, 7(5), 1995.
5. Montes-y-Gómez, López-López and Gelbukh (1999), Extraction of Document Intentions from Titles, Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99, Sweden, 1999.
6. Montes-y-Gómez, Gelbukh and López-López (2000), Comparison of Conceptual Graphs, Lecture Notes in Artificial Intelligence 1793, Springer 2000.
7. Montes-y-Gómez, Gelbukh, López-López and Baeza-Yates (2001), Flexible Comparison of Conceptual Graphs, To appear the Proc. DEXA-2001, 12th International Conference and Workshop on Database and Expert Systems Applications, Munich, Germany, September 3-7, 2001. Lecture Notes in Computer Science, Springer-Verlag.
8. Myaeng and Khoo (1994), Linguistic Processing of Text for a Large-Scale Conceptual Information Retrieval System, Lecture Notes in AI 835, Springer-Verlag 1994.
9. Sowa (1984), Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, M.A., 1984.
10. Sowa and Way (1986), Implementing a semantic interpreter using conceptual graphs, IBM Journal of Research and Development 30:1, January, 1986.